

COURS DE PROBABILITÉ IENAC

Pascal Lezaud

Ce document est disponible pour des fins pédagogiques et non commerciales. Il est protégé par le droit d'auteur.

Table des matières

Chapitre 1. Introduction et éléments du calcul des probabilités	7
1. Quelques éléments historiques	7
2. Éléments du calcul des probabilités	7
3. Dénombrements	10
4. Probabilités géométriques	15
5. Probabilités conditionnelles et indépendance	16
Chapitre 2. Variables aléatoires	25
1. Définitions	25
2. Loix fondamentales	31
3. Espérance des variables aléatoires réelles	35
4. Moments d'une variable aléatoire, inégalité de Markov	40
5. Fonction caractéristique	42
Chapitre 3. Vecteurs aléatoires	45
1. Définitions, propriétés	45
2. Loi d'une fonction de deux variables aléatoires	47
3. Espérance, covariance	50
4. Fonction caractéristique	52
5. Indépendance	53
6. Somme de variables aléatoires indépendantes	55
Chapitre 4. Espérance conditionnelle	63
1. Introduction	63
2. Conditionnement général	66
3. Loix conditionnelles	68
4. Espérance conditionnelle et projection dans $L^2(\Omega, \mathcal{A}, P)$	73
Chapitre 5. Vecteurs aléatoires gaussiens	75
1. La loi normale	75
2. Vecteurs aléatoires gaussiens	77
3. Espérances conditionnelles et variables gaussiennes	84
Chapitre 6. Convergence de suites de variables aléatoires	87
1. Convergence presque sûre	88
2. Convergence en probabilité	90
3. Convergence dans les espaces L^p	92
4. Convergence en loi	93
5. Comparaison des convergences	94

6. Lois faible et forte des grands nombres	96
7. Théorème central limite	98
8. Compléments	100
Bibliographie	103

Introduction et éléments du calcul des probabilités

1. Quelques éléments historiques

Le calcul des probabilités est né de l'étude des jeux de dés et de cartes. Le Chevalier de Méré (1607-1684), homme de lettres et philosophe est considéré comme un précurseur du calcul des probabilités modernes. Sa réputation de joueur impénitent tend à expliquer les nombreux problèmes qu'il s'est posé sur les jeux de hasard, ainsi que sa correspondance soutenue avec Fermat et Pascal.

Le calcul des probabilités s'est ensuite imposé dans la plupart des sciences physiques et humaines. Ainsi au XIX^{ème} siècle, Maxwell et Boltzmann établirent les bases de la mécanique statistique pour l'étude des gaz parfaits et Markov introduisit le concept de chaîne de Markov dans la modélisation de la langue russe. Jusqu'à la fin du XIX^{ème} siècle, les probabilités sont définies comme le rapport des cas "favorables" sur le nombre total de cas. Cependant, il est vite apparue que cette définition n'était pas satisfaisante, en particulier lorsqu'on considère un nombre infini d'événements : par exemple, dans le cas d'une suite infinie de 0 et de 1 obtenue par le lancer d'une pièce. Des problèmes se présentent également lorsque l'on considère des expériences portant sur des valeurs continues (variation de la température atmosphérique, analyse d'un signal bruité, analyse des erreurs). Le rapport du nombre de cas favorables sur le nombre total de cas doit être remplacé par le rapport d'une certaine mesure des cas favorables sur une mesure totale de l'espace des possibles considérés. Ceci se comprend bien dans le cas des probabilités géométriques, lorsque l'on manipule des surfaces dans le plan, par exemple, et illustre le lien entre la théorie des probabilités et la théorie de la mesure et de l'intégration. Or, la solution complète au problème de l'intégration est obtenue par Lebesgue en 1901 par la construction de l'intégrale qui porte son nom. La voie était ouverte à l'axiomatisation des probabilités, qui fût réalisée par A. N. Kolmogorov en 1933. Celle-ci permit alors un développement rapide et considérable du calcul des probabilités.

2. Éléments du calcul des probabilités

La théorie des probabilités a pour objet l'analyse mathématique de la notion du hasard, ou plus précisément de fournir un modèle de certains mécanismes réels. Le propos de la théorie des probabilités ne consiste donc pas à définir le hasard, mais plutôt à fournir des outils mathématiques permettant de calculer la probabilité d'apparition d'un événement "aléatoire"

(*alea* signifie dé en latin). Pour cela, il faut donc définir les notions de probabilité et d'événement. Le cas le plus simple d'une suite aléatoire peut être réalisé physiquement par une répétition illimitée de jeux de pile ou face. Considérons alors les n premiers jeux et désignons par S_n le nombre de fois que "face" est apparu. La fréquence moyenne d'apparition de "face" est S_n/n , nombre compris entre 0 et 1. On observe alors que S_n converge vers $1/2$. Ce fait, connu sous le nom de loi des grands nombres, illustre le fait qu'il est possible de dégager une certitude globale sur un phénomène non prévisible. La définition naturelle de la probabilité d'un événement, comme le rapport du nombre des cas favorables sur le nombre total des possibilités, est cohérente avec l'expérience.

2.1. Triplet de probabilités. La formalisation de Kolmogorov consiste à définir un ensemble Ω , dont chaque élément ω caractérise une possibilité théorique de réalisation du phénomène étudié. Les éléments de Ω sont encore appelés "*résultats d'expérience*" ou "*épreuves*". Un événement quelconque A est identifié à la partie de Ω composée de tous les ω qui impliquent la réalisation de A .

L'ensemble des événements, noté \mathcal{A} , doit vérifier certaines propriétés, qui dans le langage de la théorie de la mesure sont celles d'une tribu ou σ -algèbre. En effet, \mathcal{A} doit contenir *l'événement impossible* correspondant à l'ensemble vide \emptyset et *l'événement certain* associé à l'ensemble Ω . D'autre part, à tout événement A correspond *l'événement contraire*, ce dernier étant réalisé si et seulement si l'événement A ne l'est pas. L'événement contraire de A est le complémentaire de A dans Ω , il sera noté \bar{A} . Enfin, à tout couple d'événements A, B , on fait correspondre d'une part *l'événement A ou B*, correspondant à la réunion $A \cup B$, d'autre part *l'événement A et B*, correspondant à l'intersection $A \cap B$. Par définition, l'événement $A \cup B$ est réalisé si et seulement si au moins l'un des deux événements A ou B l'est, tandis que l'événement $A \cap B$ est réalisé si et seulement si les deux événements A ou B le sont.

Pour résumer, l'ensemble \mathcal{A} doit être une tribu, ou σ -algèbre de parties de Ω , c.-à-d. par définition une classe de parties de Ω contenant \emptyset, Ω et stable pour les opérations de complémentation, de réunion dénombrable et d'intersection dénombrable.

Exemple 1. On joue au dé en le lançant une fois. L'ensemble Ω peut être pris comme l'ensemble des faces du dé, $\Omega = \{1, 2, 3, 4, 5, 6\}$. Lorsqu'on lance le dé au hasard, cela revient à choisir un élément de Ω . L'ensemble des événements sera, dans ce cas, l'ensemble des parties de Ω , noté $\mathcal{P}(\Omega)$.

Exemple 2. On joue au jeu de pile ou face une infinité de fois. On comptabilise le nombre d'apparition de l'événement "face", et on s'intéresse seulement à la parité de ce nombre. Pour n lancers, désignons ce nombre par S_n . L'ensemble Ω peut être pris égal à \mathbb{N} et $\mathcal{A} = \{I, P, \emptyset, \Omega\}$, où $I = \{2n + 1, n \in \mathbb{N}\}$ et $P = \{2n, n \in \mathbb{N}\}$.

Définition 2.1. Une probabilité est une application P de l'ensemble des événements \mathcal{A} dans $[0, 1]$ telle que :

- (i) $P(\Omega) = 1$,
- (ii) (σ -additivité) Pour toute famille finie ou dénombrable $\{A_i, i \in I\}$ d'événements disjoints 2 à 2, on a $P(\bigcup_I A_i) = \sum_{i \in I} P(A_i)$.

Le triplet (Ω, \mathcal{A}, P) est appelé un espace probabilisé.

Exemple 3 (Jeu de dé). Soit $\Omega = \{1, 2, 3, 4, 5, 6\}$ muni de la tribu de ses parties. On suppose que chaque face a la même probabilité d'apparaître (événement équiprobable). L'application $P(A) = \text{Card}(A)/6$ est une probabilité, qui modélise le jeu de dé.

Exemple 4. Soit Ω la surface de la sphère de rayon R , muni de sa tribu borélienne (c.-à-d. de l'intersection des éléments de la tribu borélienne de \mathbb{R}^3 avec la sphère), soit encore des sous-ensembles de Ω suffisamment réguliers pour être mesurables. On peut définir la probabilité de trouver une mouche X sur un élément de surface S de la sphère par :

$$P(X \in S) = \frac{1}{4\pi R^2} \text{mes}(S),$$

où $\text{mes}(S)$ désigne la surface de l'élément S .

Proposition 2.2. Soient (Ω, \mathcal{A}, P) un espace de probabilité et $\{A_i, i \in I\}$ une famille finie ou dénombrable d'événements. Alors

- (i) $P(A) = 1 - P(\bar{A})$ pour tout $A \in \mathcal{A}$, en particulier $P(\emptyset) = 0$.
- (ii) Si $A_1 \subset A_2$, alors $P(A_1) \leq P(A_2)$.
- (iii) $P(\bigcup A_i) \leq \sum_{i \in I} P(A_i)$.
- (iv) $P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2)$.
- (v) $P(A_1 \cap A_2) \leq P(A_1) \leq P(A_1 \cup A_2)$.
- (vi) Si $A_i \subset A_{i+1}$ pour tout i , alors $P(\bigcup A_i) = \lim_{i \rightarrow \infty} P(A_i)$.
- (vii) Si $A_i \supset A_{i+1}$ pour tout i , alors $P(\bigcap A_i) = \lim_{i \rightarrow \infty} P(A_i)$.

Remarque 2.3. En utilisant la terminologie de la théorie de la mesure, le couple (Ω, \mathcal{A}) est un espace mesurable, et une probabilité est une mesure P sur cet espace telle que $P(\Omega) = 1$. L'étude des événements infinis conduit naturellement à imposer à \mathcal{A} d'être une tribu au lieu d'une simple algèbre de Boole.

Remarque 2.4. Par l'axiomatisation de Kolmogorov, la théorie des probabilités devient en quelque sorte une partie de la théorie de la mesure; une probabilité étant simplement une mesure de masse totale égale à l'unité. Cependant, le calcul des probabilités est plus qu'une branche de la théorie de la mesure, puisqu'il utilise des concepts nouveaux, comme les notions d'indépendance et de probabilités conditionnelles.

Exemple 5. Etant donnés deux ensembles A et B , on définit la différence symétrique de A et B par

$$A\Delta B := (A \cup B) - (A \cap B).$$

Soit (Ω, \mathcal{A}, P) un espace de probabilité et soient A et B deux événements de \mathcal{A} . Montrer que $A\Delta B \in \mathcal{A}$ et que

$$P(A\Delta B) = P(A) + P(B) - 2P(A \cap B)$$

et que

$$|P(A) - P(B)| = |P(A - B) - P(B - A)| \leq P(A\Delta B).$$

On en déduit que si $P(A\Delta B) = 0$ alors $P(A) = P(B)$.

Exemple 6. [14] Un appareil se compose de trois dispositifs, dont deux sont du premier type (A_1, A_2) et le troisième d'un second type (B). Les dispositifs A_1 et A_2 sont interchangeable, lorsque que l'un est en défaut, le second se met automatiquement en fonctionnement. Le dispositif B fonctionne indépendamment, et pour que l'appareil cesse de fonctionner il faut que les deux dispositifs A_1 et A_2 ou le dispositif B soient défectueux. Ainsi, l'événement C , correspondant à l'arrêt de l'appareil peut s'écrire comme suit :

$$C = A_1 \cap A_2 \cup B,$$

où les événements élémentaires correspondent à la défaillance des dispositifs correspondant. Exprimer la probabilité de l'événement C en fonction des probabilités des événements qui soient des unions et non des intersections des événements élémentaires.

On a

$$\begin{aligned} P(C) &= P(A_1 \cap A_2) + P(B) - P(A_1 \cap A_2 \cap B) \\ &= P(A_1) + P(A_2) - P(A_1 \cup A_2) + P(B) - P(A_1 \cap A_2 \cap B) \\ &= P(A_1) + P(A_2) - P(A_1 \cup A_2) + P(B) - \\ &\quad - (P(A_1) + P(A_2) + P(B) - P(A_1 \cup A_2)) - \\ &\quad - P(A_1 \cup B) - P(A_2 \cup B) + P(A_1 \cup A_2 \cup B) \\ &= P(A_1 \cup B) + P(A_2 \cup B) - P(A_1 \cup A_2 \cup B). \end{aligned}$$

Le lecteur intéressé par les aspects historiques et généraux de la théorie des probabilités pourra consulter [5], [13], [6], [7], [3].

Le lecteur désirant étudier, de façon approfondie, les bases mathématiques du calcul des probabilités consultera en particulier [1] et [11].

3. Dénombrements

Le raisonnement classique en probabilité consiste à attribuer, à l'événement A , la probabilité

$$P(A) = \frac{N_A}{N},$$

où N est le nombre de résultats d'expérience possibles et N_A le nombre d'épreuves favorables à l'événement A . Ceci ne sera bien sûr valable que si

chacune des épreuves est équiprobable, de probabilité de réalisation $1/N$. Un tel raisonnement ne s'appliquera pas dans le cas contraire, ou lorsque la probabilité de réalisation d'un événement élémentaire est nulle (cas d'un espace Ω infini).

3.1. Utilisation. Lorsque Ω comporte un nombre fini d'éléments, on prend très souvent comme tribu des événements, l'ensemble $\mathcal{P}(\Omega)$ des parties de Ω et comme probabilité $P(A) = \frac{\text{card}(A)}{\text{card}(\Omega)}$. Cela revient à donner la même importance à chaque épreuve, on dit alors que les événements élémentaires sont équiprobables. Dans ce cas précis, le calcul des probabilités peut s'effectuer en utilisant les dénombrements.

Exemple 7. Dans le cas du lancer de deux pièces de monnaie équilibrées, on note P l'événement "pile" et F l'événement "face". On modélisera ce jeu avec

$$\Omega = \{PP, PF, FP, FF\} = \{P, F\} \times \{P, F\}, \quad \mathcal{A} = \mathcal{P}(\Omega)$$

$$P(PP) = P(PF) = P(FP) = P(FF) = \frac{1}{4}.$$

Notons que le choix $\Omega = \{PP, H, FF\}$, où H désigne l'événement "pile, face" indépendant de l'ordre, n'est pas adapté. En effet, l'hypothèse d'équiprobabilité implique que $P(H) = 1/3$, au lieu de $1/2$ précédemment. Le choix du modèle est donc important, il doit correspondre à l'expérience réalisée et fournir les résultats obtenus. On remarquera également que dans le cas d'expériences multiples et indépendantes, le choix de Ω sera le produit cartésien des espaces d'épreuves correspondant à chacune des expériences. La tribu associée sera la tribu produit telle que définie en théorie de la mesure.

3.2. Tirages avec remise, tirages sans remise. Considérons un ensemble (ou une population) de n éléments a_1, a_2, \dots, a_n . Chaque r -uplets ordonnés, $a_{j_1}, a_{j_2}, \dots, a_{j_r}$, est appelé un *échantillon ordonné de taille r* . De façon intuitive, nous pouvons imaginer que chaque élément est sélectionné l'un après l'autre. Deux possibilités d'un tel choix existent. La première, appelée *tirage avec remise*, consiste en la sélection, à chaque tirage, d'un élément dans la population entière. Ainsi, les répétitions des éléments a_i sont possibles dans les échantillons obtenus. La seconde, appelée *tirage sans remise*, consiste à sélectionner un élément parmi les éléments de la population non encore sélectionnés par les tirages précédents. Ainsi, les répétitions ne sont pas permises, il s'ensuit que la dimension r de l'échantillon obtenu ne pourra pas dépasser la taille n de la population initiale.

Dans un tirage avec remise de r éléments parmi une population de taille n , chacun des éléments peut être choisi de n façons différentes : le nombre d'échantillons possibles est donc n^r . Dans le cas d'un tirage sans remise, nous avons n choix possibles pour sélectionner le premier élément, puis $n-1$ choix pour le second, etc., et donc $n(n-1) \cdots (n-r+1)$ choix au total. Introduisons la notation

$$(n)_r = n(n-1) \cdots (n-r+1),$$

où par convention $(n)_r = 0$ si $r > n$. Nous avons donc le résultat suivant :

Théorème 3.1. *Pour une population de taille n et un échantillon de taille r , il existe, respectivement, n^r suites ordonnées dans un tirage avec remise et $(n)_r$ dans un tirage sans remise.*

Remarquons que $(n)_n = n!$, c.-à-d. le nombre de permutations de n éléments.

Exemple 8 (Placement de boules dans des boîtes). D'après le théorème 3.1, il y a n^r façons différentes de placer r boules dans n boîtes. De nombreuses expériences peuvent se traduire en termes de placement de boules dans des boîtes. Par exemple, considérons les faces d'un dé comme des boîtes. A une expérience, consistant en r jets d'un dé, correspond un échantillon de taille r dans une population de taille 6, en particulier il y aura autant de boules dans la boîte 6 que le nombre d'apparitions de la face 6 du dé. Cette expérience donne donc lieu à 6^r résultats possibles, dont seulement 5^r satisfont à la condition qu'aucun 1 ne soit présent. Par conséquent, les tirages étant supposés équiprobables, la probabilité de n'obtenir aucun 1 dans r jets est $(5/6)^r$.

Exemple 9 (Tirage de boules avec remise). Soit une urne contenant b boules blanches et r boules rouges, telles que $b + r = N$. On effectue n tirages avec remise et on veut calculer la probabilité p_k de l'événement, noté A_k ,

“l'échantillon obtenu contient exactement k boules blanches”.

L'événement A_k peut se décomposer en la réunion des événements élémentaires, notés $A_k^{(i_1, \dots, i_k)}$:

“l'échantillon obtenu correspond aux tirages de k boules blanches aux jets i_1, \dots, i_k ”.

D'après le théorème 3.1, $\text{card}(A_k^{(i_1, \dots, i_k)}) = b^k r^{n-k}$, indépendamment de l'ordre de tirage. Le nombre d'événements élémentaires correspond au nombre de parties de k éléments dans un ensemble de n éléments, soit $\binom{n}{k}$. Ainsi

$$\text{card}(A_k) = \binom{n}{k} b^k r^{n-k}, \text{ d'où } p_k = \binom{n}{k} \left(\frac{b}{N}\right)^k \left(\frac{r}{N}\right)^{n-k}.$$

Exemple 10 (Tirage de boules sans remise). Soit une urne contenant b boules blanches et r boules rouges, telles que $b + r = N$. On effectue n tirages sans remise et on veut calculer la probabilité p_k de l'événement, noté A_k ,

“l'échantillon obtenu contient exactement k boules blanches”.

Une démarche identique à celle de l'exemple précédent conduit immédiatement à

$$p_k = \binom{n}{k} \frac{(b)_k (r)_{n-k}}{(N)_n} = \frac{\binom{b}{k} \binom{r}{n-k}}{\binom{N}{n}}.$$

3.3. Applications aux problèmes d'occupation. Considérons à nouveau l'expérience consistant à placer, au hasard, r boules dans n boîtes. Soient

r_1, r_2, \dots, r_n les nombres respectifs de boules dans les n boîtes. Chaque n -uplet vérifiant la condition

$$(3.1) \quad r_1 + r_2 + \dots + r_n = r, \quad r_k \geq 0$$

décrit une configuration possible de l'occupation des boîtes.

Notons que si les boules ne sont pas distinguables, alors deux configurations seront distinguables ssi les n -uplets (r_1, r_2, \dots, r_n) respectifs seront distincts. Nous avons le résultat suivant :

Proposition 3.2. (i) *Le nombre de configurations discernables (c.-à-d. le nombre de solutions distinctes de l'équation (3.1)) est*

$$A_{r,n} = \binom{n+r-1}{r} = \binom{n+r-1}{n-1}.$$

(ii) *Le nombre de configurations distinctes telles qu'aucune boîte ne soit vide (c.-à-d. le nombre de solution distinctes de (3.1) telles que $r_i > 0, i = 1, \dots, n$) est $\binom{r-1}{n-1}$.*

DÉMONSTRATION. (cf [6]) Représentons les boules par des étoiles et les n boîtes par n espaces entre deux barres (d'où $n+1$ barres). Ainsi $|***|**|||****|$ correspond, pour $r = 8$ boules et $n = 6$ boîtes, à la configuration $(3, 1, 0, 0, 0, 4)$. Un tel symbole nécessite de commencer et de finir par une barre, mais les $n-1$ barres restantes et les r étoiles peuvent être placées dans n'importe quel ordre. Il est alors facile de déduire, que le nombre de configurations distinctes est précisément le nombre de façons de choisir r places parmi $n+r-1$, d'où $A_{r,n}$.

La condition qu'aucune boîte ne soit vide élimine les symboles contenant des barres adjacentes (ceci impose que $r \geq n$). Ainsi, les $n-1$ barres doivent être placées dans les $r-1$ espaces laissés par les r étoiles, d'où un nombre de choix égal à $\binom{r-1}{n-1}$. \square

Proposition 3.3. *Le nombre de configurations telles que la première boîte contienne r_1 boules, la seconde r_2, \dots , la n -ième r_n est*

$$\frac{r!}{r_1!r_2! \dots r_n!}.$$

Ces coefficients sont appelés coefficients multinomiaux.

La preuve est laissée au lecteur, elle utilise les coefficients binomiaux.

Exemple 11 (Statistique de Maxwell-Boltzmann). Considérons un système composé de r particules indiscernables. En mécanique statistique, il est usuel de diviser l'espace des phases (ou espace d'état) en un grand nombre, n , de petites régions, ou cellules, de telle façon qu'à chaque particule est attribuée une cellule. Ainsi, l'état du système est décrit en termes de configurations aléatoires des r particules dans n cellules. Si les n^r configurations possibles sont équiprobables, les physiciens parleront de statistique de Maxwell-Boltzmann. Alors, la probabilité d'obtenir une configuration correspondante

aux entiers r_1, r_2, \dots, r_n , solutions de l'équation (3.1), est égale à

$$\frac{r!}{r_1!r_2!\cdots r_n!}n^{-r}.$$

Rappelons la formule suivante :

$$(a_1 + a_2 + \cdots + a_n)^r = \sum_{r_1 + \cdots + r_n = r} \frac{r!}{r_1!r_2!\cdots r_n!} a_1^{r_1} \cdots a_n^{r_n},$$

où la somme porte sur toutes les solutions de l'équation $i_1 + \cdots + i_n = r$. On vérifie ainsi que les coefficients multinomiaux forment bien un système de probabilité, puisque leur somme est égale à 1 (en prenant $a_i = 1$).

On en déduit par exemple, que la probabilité d'avoir exactement k particules dans une cellule donnée est ($r_1 = k$)

$$\begin{aligned} p_k &= \sum_{r_2 + \cdots + r_n = r - k} \frac{r!}{k!r_2!\cdots r_n!} n^{-r} = \binom{r}{k} n^{-r} \sum_{r_2 + \cdots + r_n = r - k} \frac{(r - k)!}{r_2!\cdots r_n!} \\ &= \binom{r}{k} n^{-r} (n - 1)^{r - k}. \end{aligned}$$

Le lecteur pourra essayer de retrouver ce résultat directement à partir d'un dénombrement. On démontre que p_k est maximum pour k tel que $(r - n + 1)/n \leq k \leq (r + 1)/n$ (cf [6][§II.11]).

Si $n \rightarrow \infty$ et $r \rightarrow \infty$ de telle façon que le nombre moyen $\lambda = r/n$ de particules par cellule reste constant, alors

$$p_k \rightarrow e^{-\lambda} \lambda^k / k!,$$

(distribution de Poisson).

Exemple 12 (Statistique de Bose-Einstein). La statistique de Maxwell-Boltzmann suppose les n^r configurations possibles et équiprobables. Cependant, les théories physiques modernes montrent que cette hypothèse n'est pas satisfaite par toutes les particules. En particulier, dans la statistique de Bose-Einstein les particules ne sont toujours pas distinguables, mais seules les configurations distinctes sont considérées, avec une probabilité d'apparition égale à $1/A_{r,n}$, avec $A_{r,n}$ défini à la proposition 3.2. Cette statistique s'applique entre autre aux photons.

Dans le cadre d'une statistique de Bose-Einstein, la probabilité d'avoir exactement k particules dans une cellule donnée s'exprime directement par

$$q_k = \frac{A_{r-k,n-1}}{A_{r,n}} = \frac{\binom{n+r+k-2}{r-k}}{\binom{n+r-1}{r}}.$$

On montre que pour $n > 2$, $q_0 > q_1 > \cdots$ (cf [6][§II.11]).

Si $n \rightarrow \infty$ et $r \rightarrow \infty$ de telle façon que le nombre moyen $\lambda = r/n$ de particules par cellule reste constant, alors

$$q_k \rightarrow \frac{\lambda^k}{(1 + \lambda)^{k+1}},$$

(distribution géométrique).

Exemple 13 (Statistique de Fermi-Dirac). Un autre type de comportement, connu sous le nom de statistique de Fermi-Dirac et applicable aux électrons, neutrons et protons, est basé sur les hypothèses suivantes : (1) il ne peut y avoir plus d'une particule dans une même cellule, (2) toutes les configurations distinctes, compatibles avec l'hypothèse (1), sont équiprobables. La première hypothèse implique que $r \leq n$. Une configuration est complètement décrite par la liste des n cellules non vides. Puisqu'il y a r particules, les cellules non vides peuvent être choisies de $\binom{n}{r}$ façons différentes. Ainsi, dans la statistique de Fermi-Dirac, chaque configuration a une probabilité d'apparition égale à $\binom{n}{r}^{-1}$. La probabilité d'avoir 1 particule dans une cellule donnée est

$$p = \frac{\binom{n-1}{r-1}}{\binom{n}{r}} = \frac{r}{n}.$$

4. Probabilités géométriques

Dans le cas où Ω est une partie régulière et bornée de \mathbb{R}^n , on choisira pour tribu \mathcal{A} , la tribu engendrée par l'intersection des boréliens de \mathbb{R}^n avec Ω (tribu trace). Cette tribu est engendrée par les l'intersection des pavés de \mathbb{R}^n avec Ω (cf cours de mesure intégration). Le choix naturel de cette tribu permet de pouvoir définir la mesure de tout événement A (longueur si $n = 1$, surface si $n = 2$, volume si $n = 3$, etc.). Dans ce cas, on utilise souvent la probabilité

$$P(A) = \frac{\text{mes}(A)}{\text{mes}(\Omega)},$$

correspondant à la distribution uniforme sur Ω (équiprobabilité).

Exemple 14. Soit $\Omega = [a, b]$ et soient c, d tels que $a \leq c < d \leq b$, on prendra $P([c, d]) = \frac{d-c}{b-a}$ (loi uniforme sur $[a, b]$).

Exemple 15. Donnons maintenant un exemple plus compliqué, tiré de [8]. Considérons un gaz idéal, monoatomique, composé de N molécules de masse m . On suppose que le gaz est en équilibre thermique et on notera par E son énergie totale et par $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_N$, les vitesses de chaque molécule. On cherche à calculer la probabilité que la composante v_{ix} de \mathbf{v}_i soit comprise entre α et β .

Le gaz étant supposé parfait, on négligera l'énergie potentielle des forces intermoléculaires, et l'énergie totale est alors égale à

$$2E/m = \mathbf{v}_1^2 + \dots + \mathbf{v}_N^2.$$

Il est naturel de supposer que E est proportionnel à N , soit $E = \kappa N$, où κ est une constante indépendante de N . L'état du gaz étant déterminé par les $3N$ composantes des N vecteurs \mathbf{v}_i , on prendra donc comme espace des phases, la sphère de dimension $3N$ et de rayon $R = (2E/m)^{1/2}$, notée $S_{3N}(R)$.

L'événement $\alpha \leq v_{ix} \leq \beta$ définit une zone sphérique, B , sur la sphère $S_{3N}(R)$, d'où

$$P\{\alpha \leq v_{ix} \leq \beta\} = \frac{\text{surface}(B)}{\text{surface}(S_{3N}(R))}.$$

Rappelons que les coordonnées sphériques dans \mathbb{R}^n sont

$$\begin{cases} x_1 = & r \cos \theta_1, \\ x_2 = & r \sin \theta_1 \cos \theta_2, \\ \dots & \dots \\ x_i = & r \sin \theta_1 \sin \theta_2 \cdots \sin \theta_{i-1} \cos \theta_i, \\ \dots & \dots \\ x_{n-1} = & r \sin \theta_1 \sin \theta_2 \cdots \sin \theta_{n-2} \cos \theta_{n-1}, \\ x_n = & r \sin \theta_1 \sin \theta_2 \cdots \sin \theta_{n-2} \sin \theta_{n-1}, \end{cases}$$

où $r \in [0, R]$, $\theta_i \in [0, \pi[$, $1 \leq i \leq n-2$ et $\theta_{n-1} \in [0, 2\pi[$. Le jacobien de cette transformation est

$$J(n) = r^{n-1} \sin^{n-2} \theta_1 \sin^{n-3} \theta_2 \cdots \sin \theta_{n-2}.$$

Par conséquent, l'élément de surface d'une sphère de rayon R dans \mathbb{R}^n est

$$\begin{aligned} d\sigma_n(R) &= R^{n-1} \sin^{n-2} \theta_1 \sin^{n-3} \theta_2 \cdots \sin \theta_{n-2} d\theta_1 \cdots d\theta_n \\ &= R^{n-2} (1 - x^2/R^2)^{(n-3)/2} d\sigma_{n-1}(1) dx. \end{aligned}$$

Pour obtenir la dernière ligne, nous avons fait le changement de variable $x = R \cos \theta_1$. On en déduit que :

$$P\{\alpha \leq v_{ix} \leq \beta\} = \frac{\int_{\alpha}^{\beta} (1 - mx^2/2\kappa N)^{\frac{1}{2}(3N-3)} dx}{\int_{-R}^R (1 - mx^2/2\kappa N)^{\frac{1}{2}(3N-3)} dx}.$$

Lorsque $N \rightarrow \infty$, on obtient

$$\lim_{N \rightarrow \infty} P\{\alpha \leq v_{ix} \leq \beta\} = \sqrt{\frac{3m}{4\pi\kappa}} \int_{\alpha}^{\beta} \left[\exp\left(-\frac{3mx^2}{4\kappa}\right) \right] dx.$$

On retrouve la formule de Maxwell en posant $\kappa = 3kT/2$, où k est une constante universelle et T la température absolue. La distribution, ainsi obtenue, est une loi normale de moyenne nulle et de variance $\kappa T/m$ (cf chap2).

5. Probabilités conditionnelles et indépendance

Considérons un jeu de dé dans lequel à chaque jet, chacune des six faces est équiprobable. Utilisons le modèle présenté dans l'exemple 3. Imaginons que nous lancions le dé sans le regarder, et qu'un spectateur nous dise que nous avons obtenu un chiffre pair. Etant donnée cette information, nous pouvons réévaluer la probabilité d'avoir obtenu un certain $\omega \in \Omega$. Bien entendu, si ω est impair, cette probabilité est nulle, tandis que si ω est paire celle-ci est égale à $1/3$. La façon dont nous évaluons cette probabilité, sachant que le nombre obtenu été pair, consiste à calculer $P(\{\omega\} \cap \Omega_{\text{pair}})/P(\Omega_{\text{pair}})$, où nous avons posé $\Omega_{\text{pair}} = \{2, 4, 6\}$. De façon plus générale, sur un triplet de

probabilité (Ω, \mathcal{A}, P) , si un événement B est réalisable (c.-à-d. $P(B) > 0$), on peut alors construire une nouvelle probabilité

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \quad A \in \mathcal{A},$$

appelée probabilité conditionnelle de A sachant B . Notons que $P(B|B) = 1$ et que si $A \cap B = \emptyset$, ou si $P(A \cap B) = 0$, alors $P(A|B) = 0$.

Dans certaines situations, il peut être nécessaire de conditionner par rapport à un événement de probabilité nulle. Ceci se produit en particulier, lorsque Ω est un espace non dénombrable (\mathbb{R} par exemple). A titre d'exemple, on peut chercher à estimer la probabilité que la taille d'un individu soit dans un certain intervalle, connaissant son poids. Ceci sera étudié ultérieurement.

5.1. Probabilités conditionnelles.

Définition 5.1. Soient (Ω, \mathcal{A}, P) un espace probabilisé et B un événement de \mathcal{A} tel que $P(B) > 0$. On appelle probabilité conditionnelle de l'événement $A \in \mathcal{A}$ sachant B , le nombre, noté $P(A|B)$, $P(A \cap B)/P(B)$.

Définition 5.2. Soit (Ω, \mathcal{A}, P) un espace probabilisé. Une famille d'événements $(B_i)_{i \in I}$, $I \subset \mathbb{N}$, forme un système complet d'événements si les B_i sont disjoints et $P(\bigcup_{i \in I} B_i) = \sum_{i \in I} P(B_i) = 1$.

Observons que quitte à ajouter l'événement de probabilité nulle, $N = \Omega \setminus \bigcup_{i \in I} B_i$ à la famille $(B_i)_{i \in I}$, celle-ci forme une partition de Ω .

Proposition 5.3 (formules des probabilités totales et de Bayes). Si $(B_i)_{i \in I}$ forme un système complet d'événements sur (Ω, \mathcal{A}, P) , alors pour tout $A \in \mathcal{A}$:

$$P(A) = \sum_{i \in I^*} P(A|B_i)P(B_i), \quad \text{où } I^* = \{i \in I : P(B_i) > 0\}.$$

De plus, si $P(A) > 0$, pour tout k tel que $P(B_k) > 0$, on la formule, dite de Bayes,

$$P(B_k|A) = \frac{P(A|B_k)P(B_k)}{\sum_{i \in I^*} P(A|B_i)P(B_i)}.$$

DÉMONSTRATION. La preuve est immédiate, puisque $P(A) = \sum_{i \in I^*} P(A \cap B_i)$, et pour tout k ,

$$P(B_k|A)P(A) = P(B_k \cap A) = P(A|B_k)P(B_k).$$

□

Exemple 16. Douze appareils sont en exploitation : trois d'entre eux sont fabriqués par l'usine n°1, quatre par l'usine n°2 et cinq par l'usine n°3. Ces appareils sont testés et passent l'essai avec les probabilités respectives de 0,9 pour l'usine n°1, de 0,8 pour l'usine n°2 et enfin de 0,75 pour l'usine n°3. Trouver la probabilité pour qu'un appareil choisi au hasard passe l'essai.

Comme système complet d'événements, nous pouvons choisir les hypothèses suivantes :

- : B_1 - l'appareil choisi provient de l'usine n°1,
- : B_2 - l'appareil choisi provient de l'usine n°2,
- : B_3 - l'appareil choisi provient de l'usine n°3.

Les probabilités respectives de ces hypothèses sont :

$$P(B_1) = 3/12 = 1/4,$$

$$P(B_2) = 4/12 = 1/3,$$

$$P(B_3) = 5/12.$$

Sous ces hypothèses, les probabilités conditionnelles de l'événement A : "l'appareil a passé l'essai" sont :

$$P(A|B_1) = 0,9,$$

$$P(A|B_2) = 0,8,$$

$$P(A|B_3) = 0,75.$$

En utilisant la formule des probabilités totales, on obtient $P(A) = 0,804$.

Exemple 17. Deux stations d'observation fournissent des données sur un certain système qui peut se trouver dans deux états E_1 et E_2 , tout en passant d'une manière aléatoire de l'un à l'autre. Des observations ont permis d'établir que durant 30% du temps le système se trouve dans l'état E_1 , et dans 70% dans l'état E_2 . La station d'observation n°1 fournit des données erronées dans 2% des cas, et la station d'observation n°2 dans 8% des cas. A un certain moment, la station n°1 a communiqué que le système se trouve dans l'état E_1 , et la station n°2 qu'il est dans l'état E_2 . Laquelle des communications doit être supposée exacte ?

Il est naturel de supposer vraie celle des communications dont la probabilité d'erreur est moindre. Nous allons utiliser la formule de Bayes, et à cet effet, nous choisirons comme système complet d'événements, les hypothèses suivantes :

- : B_1 - le système se trouve dans l'état E_1 ,
- : B_2 - le système se trouve dans l'état E_2 .

L'événement observé A correspond à ce que la station n°1 a transmis que le système se trouve dans l'état E_1 , et la station n°2 qu'il se trouve dans l'état E_2 .

Les probabilités des hypothèses avant l'expérience sont :

$$P(B_1) = 0,3, \quad P(B_2) = 0,7.$$

Pour que l'événement A ait lieu sous l'hypothèse B_1 , il faut que la communication transmise par la première station soit vraie et par la seconde erronée :

$$P(A|B_1) = 0,98 \cdot 0,08 = 0,0784.$$

D'une manière analogue, nous avons :

$$P(A|B_2) = 0,92 \cdot 0,02 = 0,0184.$$

En appliquant la formule de Bayes, on peut trouver la probabilité pour que l'état réel du système soit E_1 :

$$P(B_1|A) = \frac{0,3 \cdot 0,0784}{0,3 \cdot 0,0784 + 0,7 \cdot 0,0184} \approx 0,645,$$

c.-à-d. que des deux communications, celle de la première station est la plus vraisemblable.

Exemple 18 (Modèles de propagation par tirage). Pour illustrer le problème, considérons un système soumis à des défaillances. L'occurrence d'une défaillance peut être vue comme le résultat du jeu de hasard suivant : soit une urne contenant des boules rouges et des boules blanches ; à des intervalles de temps réguliers, une boule est tirée au hasard, la couleur rouge signifiant une défaillance. Mais il est concevable qu'une défaillance puisse avoir des répercussions sur le fonctionnement du système, dans le sens où elle augmente ou diminue la probabilité d'apparition d'une nouvelle défaillance. Ceci correspond à une urne dont la contenance change selon certaines règles qui dépendent du résultat du tirage. Le modèle correspondant peut être formalisé de la façon suivante : une urne contient b boules blanches et r boules rouges. Une boule est tirée au hasard. Elle est ensuite remise dans l'urne et, de plus, c boules de la couleur tirée et d boules de l'autre couleur sont ajoutées dans l'urne. Un nouveau tirage est alors effectué dans une urne contenant $r + b + c + d$ boules et la procédure est répétée. Dans ce modèle, c et d sont des entiers arbitraires pouvant être négatifs, excepté que dans ce dernier cas, le tirage peut s'arrêter au bout d'un nombre fini d'étapes. En particulier, si $c = -1$ et $d = 0$, nous obtenons le modèle de tirage sans remise.

Le résultat d'une expérience de n tirages peut se représenter par un mot de n lettres parmi l'alphabet $\{B, R\}$. L'événement "la première boule tirée est blanche" (c.-à-d. que le mot commence par B) a pour probabilité $b/(b+r)$. Si la première boule tirée est blanche, alors la probabilité (conditionnelle) d'obtenir à nouveau une boule noire est :

$$(b+c)/(b+r+c+d).$$

La probabilité (absolue) d'obtenir une séquence BB est donc

$$(5.1) \quad \frac{b}{b+r} \frac{b+c}{b+r+c+d}.$$

La probabilité d'obtenir une séquence BBB est (5.1) multipliée par $(b+2c)/(b+r+2c+2d)$. Il est possible de calculer ainsi la probabilité de tous les événements. Néanmoins, il n'existe pas de formules explicites sauf dans les cas importants suivant :

Modèle de l'urne de Polya caractérisé par $d = 0, c > 0$. L'effet d'un tirage est d'augmenter la probabilité d'apparition future de la même couleur ; nous avons alors une modélisation des effets de contagion. La probabilité d'avoir obtenu n_1 boules blanches puis n_2 boules rouges après $n = n_1 + n_2$ tirages

est donnée par

$$(5.2) \quad \frac{b(b+c)(b+2c) \cdots (b+n_1c-c) \cdots (r+c) \cdots (r+n_2c-c)}{b(b+r)(b+r+c)(b+r+2c) \cdots (b+r+nc-c)}.$$

Pour obtenir la probabilité que parmi n tirages, il y ait n_1 boules blanches et n_2 boules rouges (dans n'importe quel ordre), nous rencontrons des termes du type (5.2) arrangés dans un ordre différent. Ainsi, toutes les séquences ayant n_1 boules blanches et n_2 boules rouges sont équiprobables. Par conséquent, la probabilité, $p_{n_1, n}$ que parmi n tirages, il y ait n_1 boules blanches et n_2 boules rouges s'obtient en multipliant (5.2) par $\binom{n}{n_1}$, d'où en utilisant les coefficients binomiaux généralisés (c.-à-d. $\binom{x}{n} = (x)_n/n!$)

$$p_{n_1, n} = \frac{\binom{n_1-1+b/c}{n_1} \binom{n_2-1+r/c}{n_2}}{\binom{n-1+(b+r)/c}{n}} = \frac{\binom{-b/c}{n_1} \binom{-r/c}{n_2}}{\binom{-(b+r)/c}{n}}.$$

Modèle d'Ehrenfest de l'échange de chaleur entre deux corps isolés. Ce modèle considère deux récipients I et II contenant au total k particules. Une particule est choisie au hasard puis transportée d'un récipient dans l'autre. Cette procédure étant répétée plusieurs fois, on cherche la distribution des particules au bout de n étapes. Pour se ramener à un modèle d'urne, il suffit de dire que les particules dans le récipient I sont rouges, les autres étant blanches. Ainsi à chaque tirage, la boule choisie est remplacée par une boule d'une autre couleur, d'où $c = -1, d = 1$.

5.2. Indépendance. Dans les exemples précédents, la probabilité conditionnelle $P(A|B)$ est en général différente de $P(A)$; en quelque sorte, l'information concernant la réalisation de l'événement B modifie la probabilité de réalisation de l'événement A . Si cette information n'a aucune influence, alors $P(A|B) = P(A)$, ce que l'on traduit en disant que les événements A et B sont indépendants. Par exemple, si on jette deux fois un dé, le résultat du second jet est intuitivement indépendant du premier. Notons que la condition d'indépendance peut s'écrire $P(A \cap B) = P(A)P(B)$, relation définie même si $P(B) = 0$.

Définition 5.4. *Sur un espace probabilisé (Ω, \mathcal{A}, P) , deux événements A, B sont dit indépendants si*

$$P(A \cap B) = P(A)P(B).$$

Remarquons que si deux événements A et B sont indépendants, les tribus engendrées respectivement par A et B ($\sigma(A) = \{\emptyset, A, \bar{A}, \Omega\}$) sont indépendantes au sens où tout élément de $\sigma(A)$ est indépendant de tout élément de $\sigma(B)$. Par exemple A et \bar{B} sont indépendants puisque :

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

La définition suivante étend cette première idée intuitive de l'indépendance dans deux directions, la première pour des familles quelconques d'événements, la deuxième pour des tribus.

Définition 5.5. Soit (Ω, \mathcal{A}, P) un espace probabilisé.

- (i) Une famille quelconque d'événements $A_i \in \mathcal{A}, i \in I$, est mutuellement indépendante si pour tout $J \subset I$ fini

$$P\left(\bigcap_{j \in J} A_j\right) = \prod_{j \in J} P(A_j).$$

- (ii) Une famille quelconque de sous-tribus (ou d'algèbres) $\mathcal{A}_i \subset \mathcal{A}, i \in I$, est mutuellement indépendante si toute famille d'événements $A_i \in \mathcal{A}_i, i \in I$, est mutuellement indépendante.

Exemple 19. Prenons $\Omega = [0, 1]$ muni de sa tribu borélienne et P la mesure de Lebesgue sur $[0, 1]$. Soit, pour tout $n \in \mathbb{N}$,

$$A_n = \bigcup_{1 \leq k \leq 2^{n-1}} \left] \frac{2(k-1)}{2^n}, \frac{2k-1}{2^n} \right[.$$

La famille $(A_n)_{n \in \mathbb{N}}$ est mutuellement indépendante.

Exemple 20. Reprenons l'exemple du jet de deux dés (un rouge et un bleu). Considérons les événements

$$A = \{\text{le résultat du rouge est impair}\},$$

$$B = \{\text{le résultat du bleu est impair}\},$$

$$C = \{\text{la somme des deux dés est impaire}\}.$$

Il est aisé de vérifier que A, B, C sont indépendants deux à deux, mais ne sont pas mutuellement indépendants au sens de la définition précédente, puisque $P(A \cap B \cap C) = 0 \neq P(A)P(B)P(C)$.

5.3. Applications de l'indépendance. Ce paragraphe peut être sauté dans une première lecture. Les résultats présentés ne seront pas utilisés dans la suite du cours, bien qu'étant fondamentaux et d'un usage constant en théorie des processus stochastiques par exemple.

Définition 5.6. Soit $(\mathcal{T}_n)_{n \in \mathbb{N}}$ une famille indépendante de tribus sur (Ω, \mathcal{A}, P) . On désigne par \mathcal{A}_n la tribu engendrée par $\mathcal{T}_n, \mathcal{T}_{n+1}, \dots$ et on pose $\mathcal{A}_\infty = \bigcap_{n \in \mathbb{N}} \mathcal{A}_n$. La tribu \mathcal{A}_∞ est appelée tribu des événements terminaux ou tribu terminale.

Une interprétation intuitive de cette notion fait appel à la notion de variable aléatoire qui sera introduite ultérieurement. Considérons par exemple une expérience consistant en un nombre infini de jets de dé. Introduisons alors la variable aléatoire X_n correspondant au résultat du n -ième jet ($X_n \in \{1, 2, 3, 4, 5, 6\}$). Prenons alors pour tribu \mathcal{T}_n , la tribu engendrée par tous les événements observables au n -ième jet et donc \mathcal{A}_n sera la tribu des tous les événements observables au-delà du n -ième jet. Dans ce cas, la tribu \mathcal{A}_∞ correspond aux événements "observables à l'infini" sur la suite des jets, ou encore qui ne sont pas affectés par une modification arbitraire d'un nombre

fini d'éléments de la suite. A titre d'exemple, l'événement {pour n assez grand $X_{2n} \in A$ } est dans \mathcal{A}_∞ .

La tribu terminale vérifie la loi du tout ou rien suivante (aussi appelée loi du 0-1) :

Théorème 5.7 (Loi du 0-1). *Si \mathcal{A}_∞ est une tribu terminale, alors pour tout $A \in \mathcal{A}_\infty$, on a $P(A) = 0$, ou 1.*

Voir [11] pour une preuve rigoureuse (et relativement facile) qui consiste à prouver que la tribu terminale est indépendante d'elle-même, autrement dit $P(A \cap A) = P(A)P(A)$ pour tout $A \in \mathcal{A}_\infty$. En effet, intuitivement A , appartenant à tous les \mathcal{A}_n , est indépendant de tous les événements des tribus \mathcal{T}_m pour $m < n$. Mais la réunion de ces tribus (pour tout m) engendre une tribu contenant la tribu terminale.

Exemple 21. Soit $(A_n)_{n \in \mathbb{N}}$ une suite d'événements indépendants de (Ω, \mathcal{A}, P) , alors

$$A = \bigcap_{n \in \mathbb{N}} \bigcup_{m \geq n} A_m = \{A_n \text{ a lieu une infinité de fois}\},$$

est un événement terminal pour la suite de tribu $\mathcal{T}_n = \sigma(A_n) = \{\emptyset, \Omega, A_n, \bar{A}_n\}$. Par conséquent $P(A) = 0$ ou 1.

On abrège souvent l'expression " A_n a lieu une infinité de fois" par " A_n infiniment souvent" ou " A_n i.s.". Remarquez que $P(A_n \text{ i.s.}) = 0$ signifie que presque sûrement seulement un nombre fini d'événements A_n se réalise.

Théorème 5.8 (Lemme de Borel-Cantelli). *Soit $(A_n)_{n \in \mathbb{N}}$ une suite d'événements sur un espace probabilisé (Ω, \mathcal{A}, P) .*

- (i) *Si $\sum_{n \in \mathbb{N}} P(A_n) < \infty$, alors $P(A_n \text{ i.s.}) = 0$,*
- (ii) *si la suite $(A_n)_{n \in \mathbb{N}}$ est indépendante, la réciproque est vraie ;*
si $\sum_{n \in \mathbb{N}} P(A_n) = \infty$, alors $P(A_n \text{ i.s.}) = 1$.

DÉMONSTRATION. La partie (i) est évidente ; pour tout n

$$A = \bigcap_{n \in \mathbb{N}} \bigcup_{m \geq n} A_m \subset \bigcup_{m \geq n} A_m,$$

et donc $P(A_n \text{ i.s.}) = P(A) \leq \sum_{m \geq n} P(A_m)$ qui tend vers 0 avec n si la série converge.

La partie (ii) s'obtient en remarquant d'abord que pour tout n et tout $N \geq n$,

$$\begin{aligned} P\left(\bigcup_{n \leq m \leq N} A_m\right) &= 1 - P\left(\bigcap_{n \leq m \leq N} \bar{A}_m\right) \\ &= 1 - \prod_{n \leq m \leq N} (1 - P(A_m)). \end{aligned}$$

Comme $1 - x \leq e^{-x}$ pour tout $x \geq 0$,

$$P\left(\bigcup_{n \leq m \leq N} A_m\right) \geq 1 - \exp\left(-\sum_{n \leq m \leq N} P(A_m)\right).$$

Lorsque N tend vers l'infini, $\sum_{n \leq m \leq N} P(A_m)$ tend, pour tout n , vers l'infini par hypothèse, et donc

$$P\left(\bigcup_{n \leq m} A_m\right) = 1.$$

Il suffit de noter que par convergence monotone

$$P(A) = \lim_{n \rightarrow \infty} P\left(\bigcup_{n \leq m} A_m\right).$$

□

Exemple 22. On jette une infinité de fois une pièce équilibrée. Quelle est la probabilité de faire une infinité de fois deux piles consécutifs ? Représentons le jeu par une suite $(X_n)_{n \in \mathbb{N}}$ telle que $X_n = 1$ si on obtient pile et $X_n = 0$ dans le cas contraire. Les événements $\{X_n = 1\}$ et $\{X_n = 0\}$ sont indépendants et de probabilité $1/2$. Posons $A_n = \{X_n = X_{n+1} = 1\}$. On s'intéresse à $P(A_n \text{ i.s.})$. Les A_n ne forment pas une famille mutuellement indépendante, puisque par exemple la $n+1$ -ième variable détermine à la fois A_n et A_{n+1} . Par contre, la sous-suite $(A_{2n})_{n \in \mathbb{N}}$ forme une famille mutuellement indépendante. En outre, $P(A_{2n}) = 1/4$ pour tout n , et donc $\sum_{n \geq 1} P(A_{2n}) = \infty$. Ainsi, par le lemme de Borel-Cantelli, $P(A_{2n} \text{ i.s.}) = 1$. Comme $\{A_{2n} \text{ i.s.}\} \subset \{A_n \text{ i.s.}\}$, on conclut que $P(A_n \text{ i.s.}) = 1$.

Variables aléatoires

1. Définitions

Etant donné un espace probabilisé (Ω, \mathcal{A}, P) , à chaque épreuve $\omega \in \Omega$, on peut associer un élément $X(\omega)$ d'un certain espace E . Par exemple, si $\Omega = \{\text{pile, face}\}$, on peut définir une application $X : \Omega \rightarrow E$ par $X(\omega) = 1$ si $\omega = \text{pile}$, $X(\omega) = -1$ si $\omega = \text{face}$. Ici $E = \{-1, 1\}$ est un ensemble très simple, mais dans la réalité, il peut être plus compliqué et une famille $\mathcal{E} \subset \mathcal{P}(E)$, représentant les événements intéressants, peut être sélectionnée de la même façon que les événements de $\mathcal{A} \subset \mathcal{P}(\Omega)$. Soit C un événement de \mathcal{E} ; il est naturel de supposer que l'ensemble $\{\omega : X(\omega) \in C\}$ soit un événement de \mathcal{A} , c.-à-d. que $X^{-1}(C) \in \mathcal{A}$.

Dans la pratique, (E, \mathcal{E}) et X sont fixés, et les événements de Ω sont déterminés par X et E ; l'hypothèse minimum étant que $X^{-1}(C) \in \mathcal{A}$ pour tout $C \in \mathcal{E}$. On dit alors que X est une application mesurable de (Ω, \mathcal{A}, P) dans (E, \mathcal{E}) .

Définition 1.1. *Soient deux espaces mesurables (Ω, \mathcal{A}) et (E, \mathcal{E}) . Une application $X : (\Omega, \mathcal{A}) \rightarrow (E, \mathcal{E})$ est dite mesurable si et seulement si pour tout $C \in \mathcal{E}$, $X^{-1}(C) \in \mathcal{A}$. Dans le cas particulier où (Ω, \mathcal{A}) est muni d'une probabilité, on dit que X est une variable aléatoire à valeurs dans E .*

Elle sera appelée variable aléatoire réelle (en abrégé v.a.r.) si $(E, \mathcal{E}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$, où $\mathcal{B}(\mathbb{R})$ désigne la tribu des boréliens de \mathbb{R} . Elle sera appelée variable aléatoire discrète si E est fini ou dénombrable.

Une variable aléatoire n'est donc rien d'autre qu'une fonction mesurable sur un espace probabilisé. Par conséquent, tous les résultats concernant les fonctions mesurables s'appliquent aux variables aléatoires. En particulier, une v.a.r. étagée est de la forme

$$X(\omega) = \sum_{i \in J} x_i I_{A_i}(\omega),$$

où J est un ensemble d'indices fini, $(A_i, i \in J)$ une partition de Ω constituée d'éléments de \mathcal{A} , et x_i des nombres réels (I_{A_i} désigne la fonction indicatrice de l'ensemble A_i). Nous avons le théorème suivant (cf cours sur la théorie de la mesure)

Théorème 1.2. *Soit X une fonction de Ω sur $\bar{\mathbb{R}}^+$. Pour que X soit une v.a.r. positive définie sur (Ω, \mathcal{A}) , il est nécessaire et suffisant qu'elle soit la limite point à point d'une suite croissante de v.a.r. étagées positives.*

1.1. Loi d'une variable aléatoire - Fonction de répartition. Dans la pratique, lorsqu'on observe un phénomène aléatoire, on n'a souvent accès qu'aux valeurs $X(\omega)$, celles-ci pouvant être des vecteurs de \mathbb{R}^n . On ne connaît donc pas, en général, l'espace probabilisé (Ω, \mathcal{A}, P) . Cependant, il est souhaitable de pouvoir définir la probabilité d'un événement de la forme $\{X(\omega) \in A\}$ où A est un borélien de \mathbb{R}^n . Cela revient donc à définir une probabilité sur $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$. Considérons par exemple, une v.a.r. à valeurs dans \mathbb{R} , la tribu $\mathcal{B}(\mathbb{R})$ étant engendrée par les intervalles de la forme $] -\infty, x]$, il suffit de connaître la probabilité des événements $\{X(\omega) \leq x\}$, pour tout $x \in \mathbb{R}$ (ou simplement pour une famille dense dans \mathbb{R}). Ceci conduit aux notions suivantes.

Définition 1.3. Soit X une variable aléatoire de (Ω, \mathcal{A}, P) à valeurs dans (E, \mathcal{E}) . On appelle loi de X sous la probabilité P , la mesure image P^X sur (E, \mathcal{E}) . On notera parfois par $\mathcal{L}(X)$ la loi de X .

Rappelons que la mesure image de la probabilité P par la fonction mesurable X est définie pour tout $C \in \mathcal{E}$ par

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

Il est usuel et commode d'alléger les notations en posant pour tout $C \in \mathcal{E}$

$$P^X(C) = P(X \in C).$$

Le fait de ne plus mentionner ω dans les notations ne devra pas faire oublier que la probabilité s'applique sur un ensemble de Ω .

On remarque que toute probabilité est la loi d'une variable aléatoire (il suffit de prendre pour variable aléatoire, la fonction identité). Il s'ensuit que l'on peut considérer de façon complémentaire les variables aléatoires et leurs lois.

Définition 1.4. Soit X une variable aléatoire réelle définie sur un espace probabilisé (Ω, \mathcal{A}, P) . On appelle fonction de répartition de X , la fonction sur \mathbb{R} définie par

$$F^X(x) := P\{\omega : X(\omega) \leq x\} := P(X \leq x), \quad x \in \mathbb{R}.$$

Proposition 1.5. Une fonction de répartition F vérifie les propriétés suivantes :

- (i) $0 \leq F(x) \leq 1, \quad \forall x \in \mathbb{R}$,
- (ii) F est croissante, continue à droite avec une limite à gauche en tout point,
- (iii) $\lim_{x \rightarrow -\infty} F(x) = 0$ et $\lim_{x \rightarrow \infty} F(x) = 1$.

DÉMONSTRATION. (i) provient du fait que P est une probabilité. La croissance dans (ii) découle de la croissance des mesures (c.-à-d. $A \subset B \Rightarrow P(A) \leq P(B)$). La continuité à droite peut se déduire de la proposition I-2.2 (vii), puisque

$$\{X \leq t\} = \bigcap_{n \geq 1} \{X \leq t + 1/n\},$$

et que la croissance de F implique

$$\lim_{h \downarrow 0} F(t+h) = \lim_{n \rightarrow \infty} F(t+1/n) = F(t).$$

La limite à gauche est également une conséquence de la croissance de F .

La propriété (iii) vient encore de la proposition I-2.2 (vii) en remarquant que $\emptyset = \bigcap_{n \geq 1} \{X \leq -n\}$ et donc

$$0 = P(\emptyset) = \lim_{n \rightarrow \infty} P\{X \leq -n\} = \lim_{n \rightarrow \infty} F(-n),$$

tandis que $1 = P(\Omega) = \lim_{n \rightarrow \infty} P\{X \leq n\}$, d'après I-2.2 (vi). \square

La fonction de répartition d'une v.a.r. X caractérise la loi de X , c.-à-d. que $F^X = F^Y$ si et seulement si $P^X = P^Y$. En effet, si $F^X = F^Y$, alors P^X et P^Y coïncident sur les intervalles, donc sur l'algèbre et la tribu engendrées par les intervalles. La tribu engendrée par les intervalles étant la tribu borélienne, le résultat s'ensuit.

Proposition 1.6. *Une fonction de répartition admet au plus un nombre dénombrable de points de discontinuité.*

DÉMONSTRATION. Soit D_n l'ensemble des points de discontinuité avec un saut d'amplitude plus grande que $1/n$; en notant $F(t-)$ la limite à gauche de F en t ,

$$D_n = \{t \in \mathbb{R} : F(t) - F(t-) \geq 1/n\}.$$

Puisque $0 \leq F \leq 1$, nécessairement $\text{card}(D_n) \leq n$. L'ensemble des points de discontinuité, défini par $\bigcup_{n \geq 1} D_n$, est donc dénombrable. \square

Exemple 23. Considérons une variable aléatoire discrète à valeurs dans un ensemble E fini ou dénombrable, muni de la tribu de ses parties. Dans ce cas, la loi de X est définie par :

$$P(X = x_i) = p_i, \quad i \in I, \quad I \text{ fini ou dénombrable,} \quad x_i \neq x_j, \quad \text{si } i \neq j,$$

avec

$$p_i \geq 0 \quad \text{et} \quad \sum_i p_i = 1.$$

De plus, $P(X \in C) = \sum_{i: x_i \in C} p_i$.

Rappelons que la mesure de Dirac δ_x sur (E, \mathcal{E}) est la probabilité définie par $\delta_x(A) = 1$ si $x \in A$ et 0 si $x \notin A$. Par conséquent, la loi de la variable aléatoire discrète X est $P^X = \sum_i p_i \delta_{x_i}$. Une loi discrète est donc une combinaison finie ou dénombrable de masses de Dirac. Nous verrons ultérieurement des exemples fondamentaux de loi discrète.

Exemple 24. Soit X une v.a.r. Supposons qu'il existe une fonction f positive et mesurable par rapport à la mesure de Lebesgue, telle que pour tout borélien C , $P^X(C) = \int_C f(x) dx$. Cette fonction f est appelée la densité de la loi P^X (par rapport à la mesure de Lebesgue) et on dit également que X est de densité f . Sa fonction de répartition s'écrit

$$F^X(x) = \int_{-\infty}^x f(t) dt.$$

On en déduit que la fonction f doit vérifier la condition supplémentaire suivante :

$$\int_{-\infty}^{\infty} f(t)dt = 1.$$

Heuristiquement, la densité peut s'interpréter de la façon suivante :

$$f(x)dx = P(x \leq X < x + dx) + o(dx).$$

Des exemples fondamentaux de loi admettant une densité seront donnés ultérieurement.

1.2. Compléments. Nous présentons dans cette partie des résultats utiles, voir important d'un point de vue théorique, mais dont la lecture peut être omise sans nuire à la compréhension de la suite du cours.

1.2.1. *Théorème de Radon-Nikodym.* Nous introduisons, sans démonstration, le concept de densité d'une loi de probabilité dans un contexte plus général, en s'appuyant sur un théorème fondamental de la théorie de la mesure : le théorème de Radon-Nikodym. Avant cela, considérons une façon d'engendrer des probabilités à partir d'une fonction mesurable positive, ce que justement nous avons fait auparavant.

Proposition 1.7. *Soit f une fonction mesurable et positive définie sur un espace mesuré (E, \mathcal{E}, μ) (la mesure μ étant supposée σ -finie). On définit une fonction d'ensembles ν sur \mathcal{A} par $\nu(A) = \int_A f d\mu, A \in \mathcal{E}$. Alors ν définit une mesure. De plus, si $\mu(A) = 0$, alors $\nu(A) = 0$. Enfin, si g est une fonction positive, $\int g d\nu = \int g f d\mu$. De plus, g est ν -intégrable si et seulement si $g f$ est μ -intégrable et dans ce cas $\int g d\nu = \int g f d\mu$.*

Cette proposition dit simplement qu'à partir d'une fonction mesurable et positive, il est possible de construire une mesure sur (E, \mathcal{E}) . Dans la définition de ν , l'intégrale est prise au sens de la théorie de la mesure, c.-à-d. intégrale de la fonction $I_A f$ sur la mesure μ , où I_A est l'indicatrice de l'ensemble A . Par exemple, supposons $(E, \mathcal{E}, \mu) = (\mathbb{R}, \mathcal{B}(\mathbb{R}), \lambda)$, où λ est la mesure de Lebesgue. Soit f mesurable, positive et telle que $\int_{\mathbb{R}} f d\lambda = 1$, alors la mesure ν est une probabilité sur \mathbb{R} , qui s'annule sur les ensembles de mesure de Lebesgue nulle.

Définition 1.8. *Soient deux mesures μ et ν définies sur un espace mesurable (E, \mathcal{E}) . On dit que ν est absolument continue par rapport à μ , si tout ensemble de mesure nulle pour μ est de mesure nulle pour ν (c.-à-d. $\mu(A) = 0 \Rightarrow \nu(A) = 0$). On note alors $\nu \ll \mu$. De plus, si $\nu \ll \mu$ et $\mu \ll \nu$, on dit que ν et μ sont équivalentes.*

La proposition 1.7 fournit donc une classe de mesures absolument continues par rapport à une mesure μ . Le théorème de Radon-Nikodym montre que c'est une situation générale.

Théorème 1.9 (Théorème de Radon-Nikodym). *Si μ et ν sont deux mesures σ -finies, telles que $\nu \ll \mu$, alors il existe une fonction positive f telle que $\nu(A) = \int_A f d\mu$ pour tout ensemble mesurable A . On note $f = \frac{d\nu}{d\mu}$ et f est appelée la densité de ν par rapport à μ .*

Pour la preuve, voir par exemple [11]. Dans le cas d'une loi P admettant une densité f , cette densité n'est autre que $f = dP/d\lambda$ au sens du théorème de Radon-Nikodym, où λ est la mesure de Lebesgue. Une condition suffisante pour cela est que la loi P soit absolument continue par rapport à la mesure de Lebesgue. On notera que ce théorème est une généralisation dans le cas de la théorie de la mesure, du résultat concernant la dérivée d'une primitive d'une fonction continue. Notons cependant que la densité f n'est pas unique, il suffit de modifier sa valeur sur un ensemble négligeable pour la mesure de Lebesgue.

Supposons maintenant que la loi P^X , d'une v.a.r. X , admette un atome, c.-à-d. qu'il existe un point x tel que $P(X = x) > 0$. Cette loi n'est pas absolument continue par rapport à la mesure de Lebesgue, puisque $\lambda(\{x\}) = 0$; elle n'admet donc pas de densité. Une loi concentrée sur des atomes n'est autre qu'une loi discrète. Cependant une loi peut admettre des atomes sans être discrète. Le théorème suivant précise ce point, pour une démonstration voir par exemple [11] :

Théorème 1.10. *Soient μ et ν deux mesures σ -finies sur (E, \mathcal{E}) . Alors il existe une unique décomposition $\nu = \nu_{ac} + \nu_{\perp}$ avec $\nu_{ac} \ll \mu$ et ν_{\perp} étrangère par rapport à μ , c.-à-d. telle qu'il existe un ensemble $A \in \mathcal{E}$ avec $\nu_{\perp}(A) = 0$ et $\mu(A) = 0$. La mesure ν_{ac} est appelée la partie absolument continue de ν par rapport à μ , et ν_{\perp} la partie étrangère par rapport à μ .*

Exemple 25. Soit F une fonction de répartition. Soit $(x_n)_{n \in I}$, $I \subset \mathbb{N}$, la suite des points de discontinuité de F et $p_n = F(x_n) - F(x_n-)$ le saut correspondant. On peut poser $\mathcal{F}_d = \sum_{n \in I} p_n I_{[x_n, \infty]}$. Soit $\alpha = \lim_{t \rightarrow \infty} \mathcal{F}_d(t)$. Si $\alpha = 0$, la fonction F est continue. Sinon, $F_d = \alpha^{-1} \mathcal{F}_d$ est une fonction de répartition discrète. C'est en fait la fonction de répartition de la mesure de probabilité $\frac{1}{\alpha} \sum_{n \in I} p_n \delta_{x_n}$. Si $\alpha = 1$, alors $F = F_d$ est discrète. Sinon, $F_c = \frac{1}{1-\alpha}(F - \mathcal{F}_d)$ est une fonction de répartition continue. Ainsi $F = \alpha F_d + (1 - \alpha)F_c$ est la moyenne d'une fonction de répartition continue et d'une fonction de répartition discrète.

Notons P_c , la probabilité sur $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ de fonction de répartition F_c . En décomposant P_c suivant la mesure de Lebesgue conformément au théorème précédent, on peut écrire $P_c = \beta P_{ac} + (1 - \beta)P_{\perp}$ pour un $\beta \in [0, 1]$; P_{ac} étant une probabilité absolument continue par rapport à la mesure de Lebesgue, et P_{\perp} lui étant singulière. Notons F_{ac} et F_{\perp} les fonctions de répartition respectives. L'absolue continuité de P_{ac} assure l'existence d'une densité f , ainsi

$$F(x) = (1 - \alpha)\beta \int_{-\infty}^x f(y)dy + (1 - \alpha)(1 - \beta)F_{\perp}(x) + \alpha F_d(x).$$

La loi associée à F_d est discrète, tandis que celle associée à F_{\perp} est continue mais étrangère à la mesure de Lebesgue. Dans la pratique courante F_{\perp} sera nulle.

1.2.2. *Taux de hasard d'une variable aléatoire.* Soit une v.a.r. T admettant une densité f par rapport à la mesure de Lebesgue sur \mathbb{R}^+ . F désignera

la fonction de répartition de T , et nous poserons $\bar{F}(t) = 1 - F(t)$. On appelle *taux de hasard* de la v.a. T , la fonction

$$h(t) = \begin{cases} \frac{f(t)}{\bar{F}(t)} & \text{si } \bar{F}(t) \neq 0, \\ 0 & \text{si } \bar{F}(t) = 0. \end{cases}$$

Remarquons que la fonction f n'étant définie qu'à une équivalence près (relativement à la mesure de Lebesgue), il en est de même de h . Le terme de "taux" se justifie par le résultat suivant :

Supposons que f soit continue sur \mathbb{R}^{+} . Alors, pour tout $t > 0$ tel que $P(T > t) > 0$:*

$$h(t) = \lim_{\Delta \rightarrow 0^+} \frac{1}{\Delta} P\{t < T \leq t + \Delta | T > t\}.$$

Exemple 26. Considérons un matériel (composant électronique, moteur, etc.) pouvant se trouver dans deux états : l'état de bon fonctionnement et l'état de panne. Supposons qu'à l'instant $t = 0$, le matériel en question est dans l'état de bon fonctionnement. Soit T la v.a.r. représentant la durée de fonctionnement du matériel ; T est donc le premier instant t tel que le matériel soit panne. On appelle *fiabilité* du matériel à l'instant t , la probabilité $R(t)$ que le matériel soit en fonctionnement sur tout l'intervalle de temps $[0, t]$. Par conséquent

$$R(t) = P(T > t) = \bar{F}(t).$$

Le taux de hasard de T est dénommé *taux de défaillance* et noté $\lambda(t)$; il représente la probabilité instantannée que le matériel défaille à l'instant t , sachant qu'il a fonctionné sur tout l'intervalle $[0, t]$.

Nous avons la caractérisation fondamentale suivante du taux de hasard.

Proposition 1.11. *La v.a. T admet pour taux de hasard h , si et seulement si pour tout t positif*

$$\bar{F}(t) = P(T > t) = \exp\left(-\int_0^t h(s) ds\right).$$

Reprenons l'exemple ci-dessus, nous en déduisons que la fiabilité d'un matériel de taux de défaillance λ est :

$$R(t) = \exp\left(-\int_0^t \lambda(s) ds\right).$$

Le second corollaire de la proposition précédente est une manière d'exprimer le fait que la loi exponentielle est la seule loi avec densité qui soit sans mémoire. En effet, T a un taux de hasard constant égal à c si et seulement si T est de loi exponentielle de paramètre c .

En particulier, si la v.a. T désigne la durée de fonctionnement d'un matériel, un taux de défaillance constant signifie que le matériel ne vieillit pas. Il est couramment admis que la courbe de la fonction $\lambda(t)$ est une courbe en "baignoire", c.-à-d. que dans un première période $\lambda(t)$ décroît (période dite

de déverminage ou de rodage), puis le taux de défaillance est approximativement constant (période de vie utile), enfin dans une troisième phase le taux de défaillance est croissant (période de vieillissement ou d'usure).

Il est possible de généraliser la notion de taux de hasard à des variables aléatoires ne possédant pas de densité. On parle alors de "mesure de hasard" $dH(s) = \frac{1}{1-F(s)}dF(s)$. Pour plus de détails sur les taux de hasard et les notions de fiabilité des systèmes, le lecteur pourra consulter l'excellent ouvrage [4].

2. Lois fondamentales

2.1. Loi de Bernoulli. Une variable aléatoire X définie sur (Ω, \mathcal{A}, P) , à valeurs dans $\{0, 1\}$, suit une loi de Bernoulli, notée $\mathcal{B}(1, p)$ si

$$P(X = 1) = p = 1 - P(X = 0).$$

L'expérience du jet d'une pièce conduit naturellement à une v.a.r. de loi de Bernoulli, en posant par exemple $X = 0$ si pile et $X = 1$ si face. On aura en particulier une loi $\mathcal{B}(1, 1/2)$ si la pièce est équilibrée.

2.2. Loi binomiale. Une variable aléatoire X définie sur (Ω, \mathcal{A}, P) , à valeurs entières, suit une loi binomiale de taille $n \geq 1$ et de paramètre p , notée $\mathcal{B}(n, p)$, si

$$P(X = k) = \begin{cases} \binom{n}{k} p^k (1-p)^{n-k} & \text{si } k = 0, 1, \dots, n \\ 0 & \text{sinon} \end{cases}$$

Nous avons déjà rencontré cette loi dans le cas d'un tirage avec remise (cf exemple 1-(10)); la variable aléatoire "nombre de boules blanches tirées" suit une loi $\mathcal{B}(n, b/N)$.

Soit $(X_i)_{1 \leq i \leq n}$ une suite de variables aléatoires de loi de Bernoulli indépendantes, alors $X_1 + X_2 + \dots + X_n$ suit une loi $\mathcal{B}(n, p)$. La notion de v.a.r. indépendantes sera abordée ultérieurement; heuristiquement celle-ci signifie simplement, dans ce cas, que les événements définis par ces variables sont indépendants.

2.3. Loi de Poisson. Une variable aléatoire X définie sur (Ω, \mathcal{A}, P) , à valeurs entières, suit une loi de Poisson de paramètre $\lambda > 0$, notée $\mathcal{P}(\lambda)$, si

$$P(X = k) = e^{-\lambda} \frac{\lambda^k}{k!} \quad \text{si } k \in \mathbb{N}.$$

L'émission aléatoire d'une source radioactive peut être approximée par une loi de Poisson. En effet, on peut considérer que le nombre de particules émises par une source radioactive pendant un intervalle de temps Δt suit une loi de Poisson de paramètre $\lambda \Delta t$.

Considérons une loi binomiale $\mathcal{B}(n, p)$ et posons $b(k; n, p) = P(X = k)$. Supposons maintenant que n est suffisamment grand et p suffisamment petit

pour que le produit $\lambda = np$ soit une grandeur constante. Pour $k = 0$, nous pouvons écrire du fait que n est grand

$$b(0; n, p) = (1 - p)^n = \left(1 - \frac{\lambda}{n}\right)^n \approx e^{-\lambda}.$$

En utilisant la relation

$$b(k; n, p) = b(k - 1; n, p) \frac{\lambda - (k - 1)p}{k(1 - p)} \approx \frac{\lambda}{k} b(k - 1; n, p),$$

on obtient par récurrence, l'approximation

$$b(k; n, p) \approx e^{-\lambda} \frac{\lambda^k}{k!}.$$

On montre en fait que

$$\lim_{n \rightarrow \infty} b(k; n, \frac{\lambda}{n}) = e^{-\lambda} \frac{\lambda^k}{k!},$$

c.-à-d. que la loi de Poisson est la limite de la loi binomiale de paramètre λ/n . La loi de Poisson est également appelée loi des événements rares, en raison de la propriété d'approximation précédente. En effet, pour λ constant, lorsque n croît la probabilité de l'événement p décroît ; la loi de Poisson exprime donc la répartition binomiale pour un grand nombre d'expériences et une probabilité infime de l'événement.

Exemple 27. Plaçons n points au hasard dans l'intervalle $[0, T]$. Quelle est la probabilité que k de ces points soient dans l'intervalle $[t_1, t_2]$, où $0 < t_1 < t_2 < T$. Cet exemple peut être vu comme une répétition de n placements d'un seul point dans l'intervalle $[0, T]$. Dans ce cas, l'événement $A = \{\text{le point est dans l'intervalle } [t_1, t_2]\}$ admet pour probabilité la loi uniforme (cf ci-dessous)

$$P(A) = \frac{t_2 - t_1}{T} := p.$$

Lorsque l'on répète n fois cette expérience, on obtient un tirage avec remise, d'où

$$P\{k \text{ points sont dans l'intervalle } [t_1, t_2]\} = \binom{n}{k} p^k (1 - p)^{n-k}.$$

Quant T et n tendent vers l'infini de telle façon que le rapport $\lambda = n/T$ reste constant, alors la probabilité que k points tombent dans l'intervalle $[t_1, t_2]$ s'exprime par une loi de Poisson de paramètre $\lambda(t_2 - t_1)$.

2.4. Loi uniforme sur $[a, b]$. Une variable aléatoire X suit une loi uniforme sur $[a, b]$, $a < b$, notée $\mathcal{U}_{[a,b]}$, si sa densité par rapport à la mesure de Lebesgue sur \mathbb{R} est

$$f(x) = \frac{1}{b - a} I_{[a,b]}(x),$$

où $I_{[a,b]}$ est l'indicatrice de l'intervalle $[a, b]$.

La loi $\mathcal{U}_{[0,1]}$ est fondamentale puisque pour simuler une variable aléatoire de fonction de répartition F , il suffit de savoir simuler une variable aléatoire uniforme sur $[0, 1]$. En effet, nous avons le résultat suivant :

Proposition 2.1. *Soit F une fonction de répartition. On appelle fonction de quantile la fonction*

$$F^{\leftarrow}(u) = \inf\{x : F(x) \geq u\}, \quad u \in]0, 1[.$$

Si U est une v.a.r. de loi uniforme sur $]0, 1[$, alors $F^{\leftarrow}(U)$ est une v.a.r. de fonction de répartition F .

DÉMONSTRATION. Remarquons tout d'abord que la fonction de quantile est finie sur $]0, 1[$ et croissante. En raison de la croissance de F , si F est inversible alors la fonction de quantile est la fonction inverse. Par conséquent, sur tout intervalle de croissance stricte de F la fonction de quantile est l'inverse de F , tandis que sur un intervalle de constance de F la fonction de quantile présente un saut et en tout point de discontinuité de F , la fonction de quantile est constante (cf. figure 1). On démontre également que la fonction quantile est continue à droite et admet une limite à gauche en tout point.

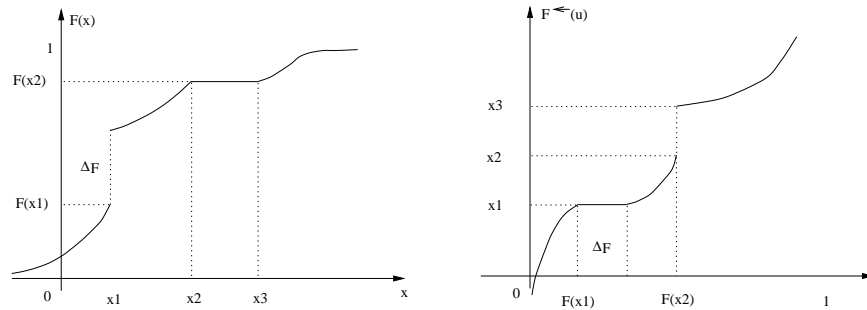


FIG. 1. Fonction de quantile

Pour prouver la proposition, observons que $\{F^{\leftarrow}(U) \leq x\} \subset \{F(x) \geq U\}$ et $\{U < F(x)\} \subset \{F^{\leftarrow}(U) \leq x\}$. Donc

$$\begin{aligned} P(F^{\leftarrow}(U) < x) &\leq P(U \leq F(x)) \\ &= F(x) \\ &= P(U < F(x)) \leq P(F^{\leftarrow}(U) \leq x). \end{aligned}$$

En particulier, puisque les fonctions de répartition sont continues à droite,

$$\begin{aligned} F(x) &\leq P(F^{\leftarrow}(U) \leq x) = \lim_{\epsilon \downarrow 0} P(F^{\leftarrow}(U) < x + \epsilon) \\ &\leq \lim_{\epsilon \downarrow 0} F(x + \epsilon) = F(x). \end{aligned}$$

Donc $P(F^{\leftarrow}(U) \leq x) = F(x)$ pour tout $x \in \mathbb{R}$, ce qui est le résultat. \square

Il existe plusieurs générateurs de nombres aléatoires uniformément distribués sur $[0, 1]$ (fonction `rand()` en C). A partir d'une séquence (x_i) de tels nombres, il est possible de générer une séquence de nombres aléatoires

(y_i) distribués selon la fonction de répartition F ; il suffit pour cela de poser $y_i = F^{\leftarrow}(x_i)$. Cette méthode est très facile à mettre en œuvre lorsque la fonction F est inversible. Dans le cas contraire, on peut utiliser certaines propriétés de F (par ex. génération de v.a.r. de loi $\mathcal{N}(0, 1)$) ou la méthode dite “du rejet”. Une bonne référence dans ce domaine est [9].

2.5. Loi exponentielle. Une variable aléatoire X définie sur (Ω, \mathcal{A}, P) et à valeurs dans \mathbb{R}^+ suit une loi exponentielle de paramètre $\lambda > 0$ si sa densité par rapport à la mesure de Lebesgue est

$$f(x) = \lambda e^{-\lambda x}, \quad x \geq 0.$$

Exemple 28. Reprenons l'exemple (27) sur la loi de Poisson, c.-à-d. considérons la variable aléatoire N_{t_0} donnant le nombre de points dans l'intervalle $[0, t_0]$, pour un $t_0 > 0$ fixé. Par conséquent

$$P(N_{t_0} = k) = e^{-\lambda t_0} \frac{(\lambda t_0)^k}{k!}.$$

Soit T_1 le premier point à droite de t_0 , et définissons la v.a.r. X comme la distance entre t_0 et T_1 . Il s'ensuit que $X(\omega) \geq 0$ pour tout ω . Ainsi la fonction de répartition de X est nulle pour tout $x < 0$. D'autre part, l'événement $X \leq x$ pour $x \geq 0$ signifie qu'il y a au moins un point dans l'intervalle $[t_0, t_0 + x]$. Donc $1 - F(x)$ est égal à la probabilité p_0 qu'il n'y ait aucun point dans l'intervalle $[t_0, t_0 + x]$, soit

$$1 - F(x) = p_0 = e^{-\lambda x}, \quad x \geq 0.$$

Ce qui est précisément, la fonction de répartition d'une v.a.r. suivant une loi exponentielle.

Signalons également que la fonction de répartition d'une loi exponentielle est inversible, il est donc très facile de générer une suite de nombres aléatoires suivant cette loi, à partir d'une suite de nombres uniformément distribués sur $[0, 1]$.

2.6. Loi normale unidimensionnelle. Une variable aléatoire X définie sur (Ω, \mathcal{A}, P) suit une loi normale de moyenne m et de variance σ^2 , notée $\mathcal{N}(m, \sigma^2)$, si sa densité par rapport à la mesure de Lebesgue est

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-m)^2}{2\sigma^2}\right), \quad x \in \mathbb{R}.$$

Nous avons déjà rencontré cette loi, appelée aussi loi de Gauss, dans le modèle du gaz parfait (cf exemple I-13), puisque les composantes du vecteur vitesse est une v.a.r. qui suit une loi $\mathcal{N}(0, \frac{2\kappa}{3m})$.

On peut prouver que si $np(1-p) \gg 1$,

$$b(k; n, p) := \binom{n}{k} p^k (1-p)^{n-k} \sim \frac{1}{\sqrt{2\pi np(1-p)}} \exp\left(-\frac{(k-np)^2}{2np(1-p)}\right),$$

pour tout k dans un voisinage de np de largeur $\sqrt{np(1-p)}$. La preuve (cf [6]) est basée sur la formule de Stirling

$$n! \sim \sqrt{2\pi n} n^{n+\frac{1}{2}} e^{-n},$$

où le signe \sim signifie que le rapport des deux termes tend vers l'unité lorsque n tend vers zéro.

L'erreur dans l'approximation normale de $b(k; n, p)$ sera d'autant plus faible que $np(1-p)$ sera grand. D'autre part, si n est grand et p petit nous pouvons approximer $b(k; np)$ par une loi de Poisson avec $\lambda = np$. Pour λ petit, seule l'approximation de Poisson doit être utilisée, mais pour λ grand, on peut utiliser les deux approximations.

2.7. Loi gamma. Une variable aléatoire X définie sur (Ω, \mathcal{A}, P) suit une loi gamma de paramètres $\nu > 0$ et $\theta > 0$, notée $\gamma(\nu, \theta)$, si sa densité par rapport à la mesure de Lebesgue sur \mathbb{R} est

$$f(x) = \frac{\theta^\nu}{\Gamma(\nu)} e^{-\theta x} x^{\nu-1} I_{[0, \infty[}(x),$$

où $\Gamma(\nu) = \int_0^\infty e^{-x} x^{\nu-1} dx$ est la fonction gamma. Rappelons que pour $x > 0$ $\Gamma(x+1) = x\Gamma(x)$, d'où pour tout n entier positif, $\Gamma(n+1) = n!$. De plus, $\Gamma(1/2) = \sqrt{\pi}$.

Pour $\nu = n$ on obtient la loi dite d'Erlang, qui donne pour $n = 1$ la loi exponentielle. Pour $\nu = n/2$ et $\theta = 1/2$ on obtient la loi dite du $\chi^2(n)$ à n degrés de liberté.

3. Espérance des variables aléatoires réelles

Dans un jeu de hasard, soit X la variable aléatoire symbolisant le gain obtenu par un joueur. Supposant que les différentes possibilités de gagner soient S_1 avec la probabilité p_1 , S_2 avec la probabilité p_2 , \dots , S_k avec la probabilité p_k , avec $p_1 + \dots + p_k = 1$. On définit alors l'espérance de gain par

$$E(X) = p_1 S_1 + \dots + p_k S_k.$$

La notion d'espérance a été initialement introduite par Huygens en 1657. Le nom d'espérance y apparaît en latin sous le nom de *expectatio*, avec l'interprétation d'être "le juste prix auquel un joueur accepterait de céder sa place dans une partie". L'interprétation de Huygens est très suggestive pour comprendre la notion d'espérance, mais elle recèle des pièges.

Soit (Ω, \mathcal{A}, P) un espace probabilisé, on associe à toute v.a.r. étagée définie sur (Ω, \mathcal{A}) le nombre réel $\sum_I x_i P(A_i)$ (notations du § 1). On appelle ce nombre l'espérance de X et on le note $E(X)$, $\int X(\omega) P(d\omega)$, ou encore $\int X dP$.

Proposition 3.1. *L'espérance $E(\cdot)$ définie sur l'espace vectoriel \mathcal{E} des v.a.r. étagées sur (Ω, \mathcal{A}) à partir de la probabilité P est l'unique forme linéaire, positive sur \mathcal{E} telle que $E(I_A) = P(A)$. Elle jouit de plus de la propriété de continuité monotone séquentielle : $X_n \downarrow X$ (resp. \uparrow) alors $E(X_n) \downarrow E(X)$ (resp. \uparrow).*

Pour une démonstration voir [11]. Etant donné un espace de probabilité (Ω, \mathcal{A}, P) et l'espérance $E(\cdot)$ qui lui est associé sur l'espace \mathcal{E} des v.a.r. étagées, il est possible, en s'appuyant notamment sur le théorème 1.2, de prolonger cette espérance à l'ensemble de toutes les v.a.r. positives sur (Ω, \mathcal{A}, P) tout en préservant les propriétés de linéarité, de croissance et de continuité monotone de cette espérance. Soit maintenant une v.a.r. X définie sur (Ω, \mathcal{A}, P) non nécessairement positive. Définissons alors les v.a.r. positives $X^+ = \sup(0, X)$ et $X^- = \sup(0, -X)$ vérifiant la relation $X = X^+ - X^-$. On dira alors que la v.a.r. est *intégrable* si $E(X^+) < \infty$ et $E(X^-) < \infty$, et on posera $E(X) = E(X^+) - E(X^-)$; on obtient ainsi un prolongement de $E(\cdot)$ à toutes les v.a.r. intégrables qui jouisse encore des propriétés de linéarité, positivité et continuité monotone.

Remarque 3.2. L'espérance d'une v.a.r. X n'est autre que l'intégrale, au sens de la théorie de la mesure, de la fonction réelle mesurable X sous la mesure P .

Récapitulons les propriétés de l'espérance d'une v.a.r. intégrable.

Proposition 3.3. *Etant donné un espace de probabilité (Ω, \mathcal{A}, P) l'espérance $E(\cdot)$ définie sur l'ensemble des v.a.r. intégrables jouit des propriétés suivantes :*

- (i) $E(X) \geq 0$ si $X \geq 0$ ou même si $P\{X < 0\} = 0$,
- (ii) $E(cX) = cE(X)$ pour toute constante finie c ,
- (iii) Si X et Y sont deux v.a.r. intégrables, alors $E(X+Y) = E(X)+E(Y)$,
- (iv) $X \leq Y \implies E(X) \leq E(Y)$,
- (v) Soit $(X_n, n \in \mathbb{N})$ une suite de v.a.r. intégrables, alors

$$X_n \uparrow X \implies E(X_n) \uparrow E(X)$$

$$X_n \downarrow X \implies E(X_n) \downarrow E(X)$$

Rappelons le lemme de Fatou-Lebesgue,

Corollaire 3.4 (Lemme de Fatou-Lebesgue). *Si $(X_n, n \in \mathbb{N})$ est une suite de v.a.r. et si Y, Z sont des v.a.r. intégrables :*

$$X_n \leq Y \implies E(\limsup_n X_n) \geq \limsup_n E(X_n)$$

$$X_n \geq Z \implies E(\liminf_n X_n) \leq \liminf_n E(X_n).$$

En particulier, si la suite $(X_n, n \in \mathbb{N})$ est convergente et s'il existe une v.a.r. intégrable U telle que $|X_n| \leq U$ pour tout n , on a : $E(\lim_n X_n) = \lim_n E(X_n)$ (convergence dominée).

La deuxième inégalité est plus couramment utilisée sous la forme suivante : soit une suite $(X_n, n \in \mathbb{N})$ de v.a.r. positives alors $E(\liminf_n X_n) \leq \liminf_n E(X_n)$. Pour une preuve voir [11] ou un cours sur la théorie de la mesure. Rappelons brièvement que $\sup_{m \geq n} X_m \downarrow \limsup_n X_n$ et $\inf_{m \geq n} X_m \uparrow$

$\liminf_n X_n$. Un bon moyen de se rappeler les inégalités du lemme de Fatou-Lebesgue consiste à mémoriser la suite d'inégalités suivantes, valable lorsqu'il existe U intégrable telle que $-U \leq X_n \leq U$,

$$E(\liminf_n X_n) \leq \liminf_n E(X_n) \leq \limsup_n E(X_n) \leq E(\limsup_n X_n).$$

Ainsi si les X_n sont positives (donc minorées par zéro) on applique les inégalités de gauche, alors qu'une majoration des X_n conduit aux inégalités de droites.

Jusqu'à présent, dans la définition de l'espérance d'une v.a.r. X , nous n'avons pas fait intervenir la loi de X mais seulement la probabilité de l'espace d'origine.

Théorème 3.5 (de transport). *Soit X une v.a.r. définie sur l'espace de probabilité (Ω, \mathcal{A}, P) et soit ϕ une fonction borélienne de \mathbb{R} dans \mathbb{R} . Alors la v.a.r. $\phi(X)$ est intégrable si et seulement si la fonction ϕ est intégrable dans $(\mathbb{R}, \mathcal{B}(\mathbb{R}), P^X)$, P^X étant la loi de X . Si tel est le cas*

$$E(\phi(X)) = \int_{\Omega} \phi \circ X(\omega) P(d\omega) = \int_{\mathbb{R}} \phi(x) P^X(dx)$$

DÉMONSTRATION. Si $\phi = I_B$ pour un borélien B de \mathbb{R} ,

$$\int I_B dP^X = P^X(B) = P(\{\omega \in \Omega : X(\omega) \in B\}) = \int_{\Omega} I_B \circ X(\omega) dP$$

et la formule est vraie dans ce cas. Si ϕ est étagée, la formule est valide par linéarité. Si ϕ est positive, soit (ϕ_n) une suite de fonctions étagées positives convergeant en croissant vers ϕ . Alors $\phi_n \circ X$ est étagée et converge simplement en croissant vers $\phi \circ X$. En utilisant le théorème de convergence monotone, à la fois pour la probabilité P et la mesure P^X , on obtient la formule pour ϕ positive. Dans le cas général, on utilise la décomposition $\phi = \phi^+ - \phi^-$. \square

Remarque 3.6. (i) Si X est une v.a.r. intégrable alors

$$E(X) = \int_{\Omega} X dP = \int_{\mathbb{R}} x P^X(dx),$$

il suffit de choisir pour ϕ la fonction identité. La dernière intégrale se note également $\int_{\mathbb{R}} x F^X(dx)$, où F^X est la fonction de répartition associée à P^X ; cette intégrale définie par une fonction monotone croissante est appelée intégrale de Stieltjes, c.-à-d. l'intégrale définie par rapport à la mesure associée à F^X (qui n'est autre ici que la loi de X).

(ii) Si $A \in \mathcal{B}(\mathbb{R})$, la fonction indicatrice I_A est mesurable. Par définition de l'intégrale et transport

$$E(I_A(X)) = \int_{\Omega} I_A(X(\omega)) dP = \int_{\mathbb{R}} I_A(x) P^X(dx) = P^X(A) = P(X \in A).$$

- (iii) Soit X un v.a.r. admettant une densité f . Soit α une fonction bijective de \mathbb{R} dans \mathbb{R} , continue et dérivable telle que $\alpha'(x) \neq 0$ pour tout x . La variable aléatoire $Y = \alpha(X)$ a pour densité

$$g(y) = \left| \frac{d\alpha^{-1}}{dy}(y) \right| f \circ \alpha^{-1}(y).$$

En effet, si ϕ est une fonction borélienne bornée, d'après le théorème de transport et la formule de changement de variables pour les intégrales de Lebesgue, on a

$$\begin{aligned} E(\phi \circ \alpha(X)) &= \int_{\mathbb{R}} \phi \circ \alpha(x) P^X(dx) \\ &= \int_{\mathbb{R}} \phi \circ \alpha(x) f(x) dx \\ &= \int_{\mathbb{R}} \phi(y) \left| \frac{d\alpha^{-1}}{dy}(y) \right| f \circ \alpha^{-1}(y) dy. \end{aligned}$$

Exemple 29 (Loi du χ^2). Soit X une v.a.r. normale standard (c.-à-d. que X suit une loi $\mathcal{N}(0, 1)$). Considérons la fonction $\alpha(y) = y^2$. Celle-ci n'est pas bijective sur \mathbb{R} , mais seulement sur \mathbb{R}^+ et \mathbb{R}^- . On cherche à déterminer pour $a > 0$, la probabilité

$$P\{X^2 \leq a\} = P\{0 \leq X \leq \sqrt{a}\} + P\{-\sqrt{a} \leq X \leq 0\} = 2P\{0 \leq X \leq \sqrt{a}\},$$

en raison de la symétrie de la distribution gaussienne. Il ne reste plus qu'à appliquer le résultat précédent, ou à procéder à un calcul direct, pour obtenir la densité de X^2

$$g(y) = \frac{1}{\sqrt{2\pi}} y^{-1/2} e^{-y/2} I_{]0, \infty[}(y).$$

C'est la densité de la loi du $\chi^2(1)$. Par exemple, les composantes des vitesses d'une molécule d'un gaz parfait sont des v.a.r. distribuées suivant la loi normale $\mathcal{N}(0, kT/m)$ (cf. 1-(15)). On en déduit que le carré de la composante v_x admet une densité de la forme

$$g_x(y) = \left(\frac{m}{kT} \right) \frac{1}{\sqrt{2\pi}} \left(\frac{ym}{kT} \right)^{-1/2} \exp\left\{ -\frac{1}{2} \left(\frac{ym}{kT} \right) \right\}.$$

Dans la pratique, la loi de X se décompose en une partie absolument continue par rapport à la mesure de Lebesgue et une partie discrète. Si P^X a une densité f par rapport à la mesure de Lebesgue, alors sous les conditions d'intégrabilité du théorème 3.5,

$$E(\phi(X)) = \int_{\mathbb{R}} \phi(x) f(x) dx.$$

Si $P^X = \sum_{n \in \mathbb{N}} p_n \delta_{x_n}$,

$$E(\phi(X)) = \sum_{n \in \mathbb{N}} p_n \phi(x_n) = \sum_{n \in \mathbb{N}} \phi(x_n) P(X = x_n).$$

Ceci montre qu'en pratique, le calcul de $E(\phi(X))$ ne nécessite pas le calcul de la loi de $\phi(X)$.

Exemple 30. Si X suit une loi binomiale $\mathcal{B}(n, p)$, on obtient

$$\begin{aligned} E(X) &= \sum_{0 \leq k \leq n} k \binom{n}{k} p^k (1-p)^{n-k} \\ &= pn \sum_{1 \leq k \leq n} \binom{n-1}{k-1} p^{k-1} (1-p)^{(n-1)-(k-1)} \\ &= pn \sum_{0 \leq k \leq n-1} \binom{n-1}{k} p^k (1-p)^{(n-1)-k} = pn. \end{aligned}$$

De la même façon, on prouve que si X suit une loi de Poisson $\mathcal{P}(\lambda)$, alors $E(X) = \lambda$.

Exemple 31. Soit X de loi exponentielle, alors

$$E(X) = \int_0^{\infty} x \lambda e^{-\lambda x} dx = \frac{1}{\lambda}.$$

Si X suit une loi de Cauchy de densité par rapport à la mesure de Lebesgue $1/[\pi(1+x^2)]$, alors X n'admet pas d'espérance. Enfin, si X suit une loi normale $\mathcal{N}(m, \sigma^2)$, on obtient (par symétrie) que $E(X) - m = 0$, le paramètre m de la loi correspond donc à l'espérance de la v.a.r.

Rappelons à présent les inégalités de Jensen et de Hölder pour des variables aléatoires; la notation $X \in L^p$, $1 \leq p < \infty$, signifiant que $E(|X|^p) < \infty$ (cf. cours sur la théorie de la mesure pour la définition des espaces L^p , $1 \leq p \leq \infty$).

Théorème 3.7. (i) (*Inégalité de Jensen*) Si ϕ est une fonction convexe sur \mathbb{R} et si X est une v.a.r. telle que X et $\phi(X)$ sont intégrables, alors

$$\phi(E(X)) \leq E(\phi(X)).$$

(ii) (*Inégalité de Hölder*) si $X \in L^p$, $Y \in L^q$, $p, q \geq 1$ et $p^{-1} + q^{-1} = 1$, alors $XY \in L^1$ et

$$E(|XY|) \leq (E(|X|^p))^{1/p} (E(|Y|^q))^{1/q}.$$

Cette inégalité dans le cas $p = q = 2$ est appelée inégalité de Cauchy-Schwarz.

(iii) L'application $p \rightarrow (E(|X|^p))^{1/p}$ est croissante.

(iv) $\|\cdot\|_p := (E|\cdot|^p)^{1/p}$ est une norme sur $L^p(\Omega, \mathcal{A}, P)$, $p \geq 1$ (si $p = \infty$, on prendra la norme essentielle supérieure).

L'inégalité de Jensen est le plus souvent appliquée pour les fonctions $\phi(x) = |x|, x^2$ et $1/x$ lorsque $x > 0$. En particulier, une v.a.r., dont le carré est intégrable, est intégrable, et si X est à valeurs strictement positives,

$$E\left(\frac{1}{X}\right) \geq \frac{1}{E(X)}.$$

Exemple 32. Soit $Z \geq 0$ une v.a.r. telle que $EZ^2 < \infty$. Appliquant l'inégalité de Cauchy–Schwarz à la v.a.r $ZI_{Z>0}$, on obtient

$$(EZ)^2 = [E(ZI_{Z>0})]^2 \leq E(Z^2)E(I_{Z>0}) = E(Z^2)P\{Z > 0\}$$

d'où

$$P\{Z > 0\} \geq (EZ)^2/E(Z^2).$$

Soit maintenant une famille finie d'événements $(A_i; i \in I)$. Appliquons l'inégalité précédente à la variable $Z(\omega) = \sum_{i \in I} I_{A_i}(\omega)$ qui donne le nombre d'événements A_i réalisés. On obtient :

$$EZ = \sum_i P(A_i) := \mu \quad E(Z^2) = \sum_{i,j} P(A_i \cap A_j) := \sigma^2$$

d'où

$$\frac{\mu^2}{\sigma^2} \leq P\left(\bigcup A_i\right) \leq \mu,$$

l'inégalité de droite provenant de la propriété d'additivité de la probabilité.

Ce résultats permet d'obtenir des bornes pour la distribution du maximum d'une famille finie de v.a.r. X_i . En effet, soit b un réel fixé, il suffit de poser $A_i = \{X_i \geq b\}$, car $\cup A_i = \{\max_i X_i \geq b\}$.

La propriété (iii) implique que si $X \in L^q$, alors $X \in L^p$ pour tout $p \leq q$. En effet, si $X \in L^q$ alors $(E(|X|^p))^{1/p} \leq (E(|X|^q))^{1/q} < \infty$.

4. Moments d'une variable aléatoire, inégalité de Markov

Définition 4.1. Soit X une v.a.r. dont le carré est intégrable. On appelle variance de X , ou de sa loi P^X , la quantité, notée $\text{Var}(X)$

$$\text{Var}(X) = E((X - EX)^2).$$

La racine $\sqrt{\text{Var}(X)}$ est appelée l'écart-type, parfois notée $\sigma(X)$.

La variance de X s'exprime également par la relation

$$\text{Var}(X) = EX^2 - (EX)^2.$$

Exemple 33. (i) Si $\text{Var}(X) = 0$, alors X est presque sûrement constante et égale à sa moyenne EX . La variance d'une v.a.r. mesure donc la concentration de X près de sa moyenne. Donc, plus la variance est grande, plus la v.a. est dispersée, c.-à-d. peut prendre avec forte probabilité des valeurs éloignées de sa moyenne. Une autre manière de mesurer cette concentration serait de calculer la quantité $E(|X - EX|)$.

(ii) Si X suit une loi de Bernoulli $\mathcal{B}(n, p)$, sa variance est $np(1 - p)$.

(iii) Si X suit une loi $\mathcal{N}(m, \sigma^2)$, sa variance est σ^2 .

Une généralisation de la définition de la variance conduit à la notion de moment centré d'ordre p . Si X est une v.a.r. dont la puissance p -ième est intégrable, alors son moment d'ordre p est $E(X^p)$ et son moment centré d'ordre p est

$$\mu_p = E[(X - EX)^p].$$

Il peut être parfois plus commode de calculer les moments d'ordre p à partir de la fonction de répartition.

Proposition 4.2. *Soit X une v.a.r. positive, de fonction de répartition F . Alors, pour tout $0 < p < \infty$*

$$E(X^p) = p \int_0^\infty t^{p-1} P\{X > t\} dt = p \int_0^\infty t^{p-1} [1 - F(t)] dt.$$

De plus, $E(X) < \infty$ si et seulement si pour un ou tout $\epsilon > 0$,

$$\sum_n P\{X > \epsilon n\} < \infty.$$

DÉMONSTRATION. D'après le théorème de Fubini (cf. cours de théorie de la mesure)

$$\begin{aligned} p \int_0^\infty t^{p-1} P\{X > t\} dt &= p \int_0^\infty t^{p-1} E(I_{]t, \infty[}(X)) dt \\ &= E\left(p \int_0^X t^{p-1} dt\right) \\ &= E(X^p). \end{aligned}$$

Pour la seconde partie, prenons $p = 1$ et notons que

$$\sum_n P\{X > n + 1\} \leq \int_0^\infty P\{X > t\} dt \leq \sum_n P\{X > n\},$$

en découpant l'intervalle $[0, \infty[$ suivant les intervalles $[n, n+1[$. Pour conclure, il suffit de remplacer X par X/ϵ . \square

Comme cas particulier souvent employé, citons le calcul de l'espérance d'une v.a.r. positive

$$EX = \int_0^\infty (1 - F(t)) dt.$$

Dans le cas d'une variable discrète, la formule précédente devient

$$EX = \sum_{k=0}^\infty P\{X > k\}.$$

Terminons cette partie par l'énoncée de l'inégalité de Markov, inégalité essentielle dans l'étude des v.a.r.

Proposition 4.3 (Inégalité de Markov). *Soit X une v.a.r. intégrable. Alors, pour tout $t > 0$,*

$$P\{X \geq t\} \leq \frac{E(X^+)}{t} \leq \frac{E|X|}{t}.$$

DÉMONSTRATION. Observons tout d'abord que

$$I_{[t, \infty[}(X) \leq \frac{X}{t} I_{[t, \infty[}(X) \leq \frac{X^+}{t} \leq \frac{|X|}{t}.$$

Il suffit alors d'intégrer cette suite d'inégalités par rapport à la probabilité P . \square

Exemple 34. (i) Si $X \in L^p$, alors pour tout $t > 0$

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

puisque $\{X \geq t\} \subset \{|X|^p \geq t^p\}$.

(ii) Si $X \in L^2$, l'inégalité de Markov implique l'inégalité dite de Tchebycheff,

$$P\{|X - EX| \geq t\} \leq \frac{\text{Var}(X)}{t^2}, \quad t > 0.$$

(iii) Si maintenant $X \in L^p$ pour tout $p > 0$, mais $E(e^{tX}) < \infty$ pour $t > 0$, éventuellement assez petit, alors

$$P\{X \geq t\} \leq \inf_{\lambda > 0} e^{-\lambda t} E(e^{\lambda X}),$$

puisque $\{X \geq t\} \subset \{e^{\lambda X} \geq e^{\lambda t}\}$ pour tout $\lambda > 0$. Cette inégalité, dénommée inégalité de Cramér ou Chernoff, est d'un usage fréquent dans l'étude des sommes de variables aléatoires indépendantes.

5. Fonction caractéristique

Pour l'étude de la loi asymptotique d'une somme de v.a.r., A. Liapounov, en 1900, a élaboré une méthode spéciale dite des fonctions caractéristiques. Par la suite cette méthode a acquis une importance par elle-même, car comme son nom l'indique ces fonctions caractérisent les lois associées.

Définition 5.1. Soit X une v.a.r. sur (Ω, \mathcal{A}, P) à valeurs dans \mathbb{R} . On appelle fonction caractéristique de X ou de la loi de X la fonction, notée $\varphi^X(t)$, à valeurs complexes

$$\varphi^X(t) = E(e^{itX}) = \int_{\mathbb{R}} e^{itx} P^X(dx).$$

La fonction caractéristique est à valeurs complexes, de module majorée par 1 (d'après l'inégalité de Jensen appliquée à la fonction $|\cdot|$) et $\varphi^X(0) = 1$.

Si la loi de X a une densité f par rapport à la mesure de Lebesgue sur \mathbb{R} , alors

$$\varphi^X(t) = \int_{\mathbb{R}} e^{itx} f(x) dx$$

est aussi appelée la transformée de Fourier de la fonction f . L'importance de la transformée de Fourier et les propriétés très particulières des fonctions complexes analytiques tendent à indiquer que la notion de fonction caractéristique est bien adaptée. En particulier, on démontre le théorème suivant :

Théorème 5.2. *Si X et Y sont deux v.a.r. de lois respectives P^X et P^Y telles que $\varphi^X = \varphi^Y$, alors $P^X = P^Y$.*

Attention, ce résultat ne veut pas dire que $X = Y$, puisque les v.a.r X et Y peuvent être définies sur deux espaces de probabilités différents.

Exemple 35. (i) Si $X = a$ p.s., c.-à-d. que $P^X = \delta_a$ où $a \in \mathbb{R}$ est une constante, alors $\varphi^X(t) = e^{iat}$.

(ii) Si X suit une loi normale $\mathcal{N}(m, \sigma^2)$, alors

$$\varphi^X(t) = e^{itm - \sigma^2 t^2 / 2}.$$

(iii) Si X est exponentielle de densité $\lambda e^{-\lambda x}$, alors

$$\varphi^X(t) = \int_0^\infty \lambda e^{(it-\lambda)x} dx = \frac{\lambda}{\lambda - it}.$$

(iv) Si X suit une loi de Poisson de paramètre λ , alors

$$\varphi^X(t) = \sum_{n \geq 0} e^{itn} e^{-\lambda} \frac{\lambda^n}{n!} = e^{-\lambda} \sum_{n \geq 0} \frac{(e^{it}\lambda)^n}{n!} = \exp(\lambda(e^{it} - 1)).$$

Lorsque X est une v.a.r. admettant des moments de tout ordre, notons que

$$e^{itX} = \sum_{n \in \mathbb{N}} \frac{(it)^n}{n!} X^n,$$

d'où en intégrant terme à terme (sous des conditions permettant l'inversion du signe somme et de l'intégrale),

$$\varphi^X(t) = \sum_{n \in \mathbb{N}} \frac{(it)^n}{n!} E(X^n).$$

La formule de Taylor appliquée à la fonction caractéristique montre que les moments de la variable aléatoire X sont proportionnels aux dérivées de φ^X . Le résultat rigoureux est le suivant :

Proposition 5.3. *Soit X une v.a.r. de fonction caractéristique $\varphi = \varphi^X$ et de loi P^X .*

(i) *Si $E(|X|^n) < \infty$, alors φ est n -fois dérivable, et sa dérivée k -ième ($k \leq n$) est*

$$\varphi^{(k)}(t) = i^k \int x^k e^{itx} P^X(dx) = i^k E(X^k e^{itX}).$$

En particulier, $\varphi^{(k)}(0) = i^k E(X^k)$.

(ii) *Réciproquement, si φ est k -fois dérivable en 0, alors X admet des moments d'ordre plus petit ou égal à $2n \leq k$.*

DÉMONSTRATION. La lecture de la preuve peut être omise. (i) Montrons tout d'abord l'inégalité suivante, pour tout $t \in \mathbb{R}$

$$\left| e^{it} - 1 - \frac{it}{1!} - \dots - \frac{(it)^{n-1}}{(n-1)!} \right| \leq \frac{|t|^n}{n!}.$$

En effet, la fonction $f_1(t) = i \int_0^t e^{ix} dx = e^{it} - 1$ est de module plus petit que $|t|$. Par récurrence $f_n(t) = i \int_0^t f_{n-1}(x) dx$ est de module plus petit que $|t|^n/n!$.

Démontrons maintenant que φ est dérivable en tout point de \mathbb{R} lorsque $E(|X|) < \infty$. Pour tout $h \neq 0$,

$$\frac{\varphi(t+h) - \varphi(t)}{h} = \int e^{itx} \frac{e^{ihx} - 1}{h} P^X(dx).$$

D'après l'inégalité précédente pour $n = 1$, on obtient

$$\left| e^{itx} \frac{e^{ihx} - 1}{h} \right| \leq |x|$$

qui est intégrable pour P^X , par hypothèse, et ceci indépendamment de h . D'après le théorème de convergence dominée

$$\varphi'(t) = \lim_{h \rightarrow 0} \int e^{itx} \frac{e^{ihx} - 1}{h} P^X(dx) = \int ix e^{itx} P^X(dx) = iE(X e^{itX}).$$

Les dérivées d'ordre supérieur se calculent de la même façon.

(ii) D'après l'inégalité de Hölder, il suffit de montrer l'existence des moments d'ordre pair inférieur à $2n$ pour assurer celle des moments d'ordre inférieur à $2n$. Supposons prouvé l'existence du moment d'ordre $2k$, ce qui est évident pour $k = 0$. Alors d'après (i),

$$\varphi^{(2k)}(0) = (\mathcal{R}e\varphi)^{(2k)}(0) = (-1)^k E(X^{2k}),$$

puisque la partie imaginaire de φ est impaire. Il s'ensuit que

$$\begin{aligned} (-1)^{k+1} \varphi^{(2k+2)}(0) &= \lim_{h \rightarrow 0} \frac{(\mathcal{R}e\varphi)^{(2k)}(h) - (\mathcal{R}e\varphi)^{(2k)}(0)}{h} \\ &= \lim_{h \rightarrow 0} \int x^{2k} \frac{1 - \cos(hx)}{2h^2} P^X(dx). \end{aligned}$$

Utilisons le lemme de Fatou pour conclure que

$$\int x^{2k+2} P^X(dx) \leq \varphi^{(2k+2)}(0) < \infty.$$

□

Vecteurs aléatoires

Nous avons défini, en introduction au chapitre 2, la notion de variable aléatoire à valeur dans un espace E mesurable. Nous pouvons donc considérer le cas où $E = \mathbb{R}^d$ avec $d \geq 2$ muni de sa tribu borélienne engendrée par les pavés de \mathbb{R}^d (produit cartésien d'intervalles ouverts). Une telle variable est appelée, très logiquement, un vecteur aléatoire. L'utilisation des vecteurs aléatoires est courante et l'objectif de ce chapitre consistera à en présenter les principales propriétés.

Nous introduirons également dans ce chapitre la notion fondamentale d'indépendance, extension de celle déjà introduite au chapitre 1.

1. Définitions, propriétés

Définition 1.1. Soit (Ω, \mathcal{A}, P) un espace probabilisé. On appelle vecteur aléatoire une variable aléatoire à valeurs dans \mathbb{R}^d muni de sa tribu borélienne.

On montre facilement que le vecteur $X = (X_1, \dots, X_d)$ est un vecteur aléatoire si et seulement si ses composantes sont des variables aléatoires réelles.

Définition 1.2. On appelle fonction de répartition de X (ou de la loi de X) la fonction

$$t = (t_1, \dots, t_d) \in \mathbb{R}^d \mapsto F^X(t) = P\{X_1 \leq t_1, \dots, X_d \leq t_d\}.$$

La loi de la variable aléatoire X_i est appelée la i -ème loi marginale de $X = (X_1, \dots, X_d)$. Elle est donnée par

$$F^{X_i}(t_i) = \lim_{t_1, \dots, t_{i-1}, t_{i+1}, \dots, t_d \rightarrow \infty} F^X(t).$$

Exemple 36 (Couple de variables aléatoires réelles). Lorsque $d = 2$, un vecteur aléatoire n'est autre qu'un couple (X, Y) de v.a.r. La loi du couple (X, Y) permet de caractériser les liaisons et influences mutuelles des deux caractères de l'expérience correspondante. On note

$$P^{(X,Y)}(A) = P\{(X, Y) \in A\}, \quad A \subset \mathbb{R}^2,$$

la loi du couple. En particulier, si $A = A_1 \times A_2$, on pourra noter la quantité précédente

$$P^{(X,Y)}(A) = P\{X \in A_1, Y \in A_2\}.$$

Dans le cas où les v.a.r. X et Y sont discrètes, la loi du couple (X, Y) est alors définie par les coefficients p_{ij} tels que

$$P\{X = x_i, Y = y_j\} := p_{ij}, \quad (i, j) \in J \text{ fini ou dénombrable.}$$

On aura évidemment la relation $\sum_{(i,j) \in J} p_{ij} = 1$ et

$$P^{(X,Y)}(A) = \sum_{(i,j):(x_i,y_j) \in A} p_{ij}.$$

On peut aussi écrire que $P^{(X,Y)} = \sum_{(i,j) \in J} p_{ij} \delta_{(x_i,y_j)}$, où $\delta_{(i,j)}$ est la mesure de Dirac au point (x_i, y_j) .

Lorsque le couple aléatoire (X, Y) admet une densité $f(x, y)$ par rapport à la mesure de Lebesgue sur \mathbb{R}^2 , c.-à-d. que

$$F^{(X,Y)}(t_1, t_2) = \int_{-\infty}^{t_1} \int_{-\infty}^{t_2} f(x, y) dx dy,$$

alors la loi du couple s'exprime par

$$P^{(X,Y)}(A) = \int_A f(x, y) dx dy,$$

et les lois marginales de X et Y admettent respectivement une densité

$$f^X(x) = \int_{\mathbb{R}} f(x, y) dy, \quad f^Y(y) = \int_{\mathbb{R}} f(x, y) dx.$$

En particulier,

$$P\{a_1 < X \leq b_1, a_2 < Y \leq b_2\} = F^{(X,Y)}(b_1, b_2) - F^{(X,Y)}(a_1, b_2) \\ - F^{(X,Y)}(b_1, a_2) + F^{(X,Y)}(a_1, a_2),$$

et

$$f(x, y) = \lim_{\substack{\Delta x \rightarrow 0 \\ \Delta y \rightarrow 0}} \frac{P\{x < X \leq x + \Delta x, y < Y \leq y + \Delta y\}}{\Delta x \Delta y} = \frac{\partial^2 F^{(X,Y)}(x, y)}{\partial x \partial y}.$$

Heuristiquement, $f(x, y) dx dy$ est la probabilité que le point de coordonnées (X, Y) se trouve dans le rectangle élémentaire de côtés dx, dy adjacent au point (x, y) . Cette probabilité se mesure par le volume du parallélépipède élémentaire ayant pour base le rectangle élémentaire $dx dy$ et limité supérieurement par la surface $z = f(x, y)$. On en déduit que d'un point de vue géométrique, la probabilité que le point (X, Y) appartienne à un domaine D est le volume du cylindre C , limité supérieurement par la surface $z = f(x, y)$ et dont la base est le domaine D (faire un dessin). Ces remarques s'étendent sans difficulté au cas d'un vecteur aléatoire dans \mathbb{R}^d .

Notons que si $f(x, y) = h(x)g(y)$ avec $\int h(x) dx = 1$, alors $f^X = h$ et $f^Y = g$ et $P\{(X, Y) \in A_1 \times A_2\} = P\{X \in A_1\}P\{Y \in A_2\}$. Selon la terminologie du chap. 2, les événements $\{X \in A_1\}$ et $\{Y \in A_2\}$ sont indépendants et ceci pour tout A_1 et A_2 . Nous étudierons ceci plus en détail au §5.

On peut aussi définir des couples aléatoires mixtes tels que X est une variable aléatoire discrète et Y une v.a.r. admettant une densité.

Remarque 1.3. Il ressort des définitions précédentes que la loi d'un vecteur $X = (X_1, \dots, X_d)$ détermine chacune des lois marginales (lois des X_i). L'exemple suivant, tiré de [1], montre que la réciproque est fautive en générale.

Exemple 37. Supposons que le couple (X, Y) admette pour loi

$$P^{(X,Y)} = \frac{1}{4}\delta_{(-1,0)} + \frac{1}{4}\delta_{(0,1)} + \frac{1}{4}\delta_{(0,-1)} + \frac{1}{4}\delta_{(1,0)}.$$

Autrement dit, la mesure est concentrée dans \mathbb{R}^2 sur les points $(-1, 0)$, $(0, 1)$, $(0, -1)$, $(1, 0)$. Les lois marginales P^X et P^Y sont égales et données par

$$P^X = P^Y = \frac{1}{4}\delta_{-1} + \frac{1}{2}\delta_0 + \frac{1}{4}\delta_1.$$

On peut cependant construire un second vecteur aléatoire (X_1, Y_1) ayant les mêmes lois marginales dont les probabilités sont données par le tableau suivant :

	X			
	-1	1/16	1/8	1/16
Y	0	1/8	1/4	1/8
	1	1/16	1/8	1/16

2. Loi d'une fonction de deux variables aléatoires

Soit X un vecteur aléatoire à valeurs dans \mathbb{R}^d , admettant une densité f . Soit h une bijection de \mathbb{R}^d admettant une dérivée continue et de jacobien $J_h(x) \neq 0$ pour tout x . Rappelons que le jacobien de $h = (h_1, \dots, h_d)$ est le déterminant de la matrice carrée $(\partial h_i / \partial x_j)_{1 \leq i, j \leq d}$. Le vecteur $Y = h(X)$ admet alors pour densité

$$g(y) = |J_{h^{-1}}(y)| f(h^{-1}(y)).$$

Ceci se démontre de la même façon que dans le cas d'une variable aléatoire réelle, à partir de la formule du changement de variable dans une intégrale multiple. En particulier, si $d = 2$ et si la transformation est linéaire

$$X_1 = a_{11}Y_1 + a_{12}Y_2, \quad X_2 = a_{21}Y_1 + a_{22}Y_2,$$

avec un déterminant $\Delta = a_{11}a_{22} - a_{12}a_{21} > 0$, alors la densité du couple (Y_1, Y_2) s'exprime en fonction de la densité du couple (X_1, X_2) par

$$f^{(Y_1, Y_2)}(y_1, y_2) = f^{(X_1, X_2)}(a_{11}y_1 + a_{12}y_2, a_{21}y_1 + a_{22}y_2)\Delta.$$

Exemple 38. Soit le couple (X, Y) de densité :

$$f(x, y) = \frac{1}{2\pi} \exp\left(-\frac{1}{2}(x^2 + y^2)\right), \quad (x, y) \in \mathbb{R}^2.$$

Effectuons le changement de variable $(X, Y) \rightarrow (R, \Theta)$ avec :

$$X = R \cos \Theta, \quad Y = R \sin \Theta.$$

Cette transformation est une bijection de \mathbb{R}^{*2} dans $\mathbb{R}^{*+} \times [0, 2\pi[$. La matrice jacobienne correspondante est :

$$\begin{pmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \theta} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \theta} \end{pmatrix} = \begin{pmatrix} \cos \theta & -r \sin \theta \\ \sin \theta & r \cos \theta \end{pmatrix}$$

d'où $|J| = r$. Le couple (R, Θ) possède donc la densité de probabilité suivante :

$$g(r, \theta) = \frac{r}{2\pi} e^{-\frac{1}{2}r^2}, \quad (r, \theta) \in \mathbb{R}^{*+} \times [0, 2\pi[.$$

Nous en déduisons que le vecteur (X, Y) (vecteur gaussien) est un vecteur de direction choisie uniformément sur $[0, 2\pi[$ et de longueur R de loi de Rayleigh.

En particulier, soient U_1 et U_2 deux v.a.r. uniformément distribuées sur $[0, 1]$ telle que la densité jointe du couple (U_1, U_2) soit le produit des densités respectives de U_1 et U_2 (c.-à-d. que U_1 et U_2 sont indépendantes selon la définition du § 5). Posons

$$R = \sqrt{-2 \log U_1}, \quad \Theta = 2\pi U_2.$$

On vérifie facilement que Θ est distribuée uniformément sur $[0, 2\pi]$ et que R suit une loi de Rayleigh. Par conséquent, le couple (X, Y) définie par

$$X = R \cos \Theta, \quad Y = R \sin \Theta$$

admet pour densité $f(x, y)$, c.-à-d. que X et Y sont deux v.a.r. normales $\mathcal{N}(0, 1)$ indépendantes. Nous avons donc un moyen de simuler deux réalisations indépendantes d'une v.a.r. normale centrée réduite à partir de deux réalisations indépendantes d'une v.a.r. uniforme sur $[0, 1]$.

Considérons maintenant un couple (X, Y) de v.a.r. de densité $f(x, y)$. Soit φ une fonction de \mathbb{R}^2 dans \mathbb{R} . On souhaite calculer la densité de probabilité de la v.a.r. $Z = \varphi(X, Y)$. Pour cela rappelons le théorème de Fubini, dont la démonstration figure dans les cours sur la théorie de la mesure.

Théorème 2.1 (Fubini). *Soient $(\Omega_1, \mathcal{A}_1, \mu_1)$ et $(\Omega_2, \mathcal{A}_2, \mu_2)$ deux espaces mesurés pour des mesures μ_1 et μ_2 σ -finies. Considérons l'espace produit $\Omega = \Omega_1 \times \Omega_2$ muni de la tribu produit $\mathcal{A} = \mathcal{A}_1 \otimes \mathcal{A}_2$ et de la mesure produit $\mu = \mu_1 \otimes \mu_2$ (celle-ci étant définie par $\mu_1 \otimes \mu_2(A_1 \times A_2) = \mu_1(A_1)\mu_2(A_2)$). Soit f une fonction réelle, définie sur Ω , \mathcal{A} -mesurable et μ -intégrable. Alors,*

$$\begin{aligned} \int_{\Omega} f d\mu &= \int_{\Omega_1} \left(\int_{\Omega_2} f(\omega_1, \omega_2) d\mu_2(\omega_2) \right) d\mu_1(\omega_1) \\ &= \int_{\Omega_2} \left(\int_{\Omega_1} f(\omega_1, \omega_2) d\mu_1(\omega_1) \right) d\mu_2(\omega_2). \end{aligned}$$

Ce théorème signifie que, sous l'hypothèse que f soit μ -intégrable, on peut permuter les intégrales pour évaluer $\int f d\mu$. Dans la pratique, pour vérifier que f est μ -intégrable, on évalue $\int |f| d\mu$ par permutation des intégrales.

Calculons la fonction de répartition de Z :

$$F^Z(z) = P\{Z \leq z\} = P\{\varphi(X, Y) \leq z\}.$$

La fonction φ est représentée par une surface (cf. figure 1)

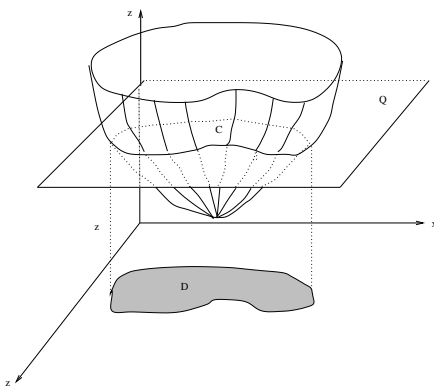


FIG. 1. Fonction de deux variables

Soit Q le plan parallèle au plan xOy situé à une distance z de celui-ci. Ce plan coupera la surface $z = \varphi(x, y)$ suivant une certaine courbe C . Projetons la courbe C sur le plan xOy . Cette projection dont l'équation est $\varphi(x, y) = z$ divisera le plan en deux domaines : pour l'un de ces domaines la cote de la surface examinée sera inférieure et pour l'autre supérieure à z . Désignons par D le domaine où cette hauteur est inférieure à z . Pour que $Z \leq z$, il faut que le point aléatoire (X, Y) se trouve dans le domaine D ; par conséquent

$$F^Z(z) = \int_D f(x, y) dx dy.$$

Exemple 39. Considérons le cas $\varphi(X, Y) = XY$. Pour une certaine valeur z , la courbe d'équation $xy = z$, dans le plan xOy est une hyperbole d'asymptotes les axes de coordonnées (il y a deux branches). Le domaine D est celui contenant le point O (cf. figure 2).

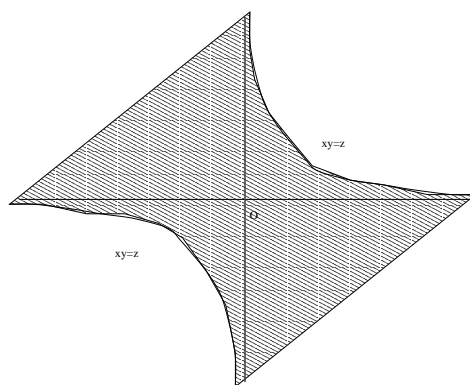


FIG. 2

La fonction de répartition de Z s'écrit :

$$F^Z(z) = \int_{-\infty}^0 \int_{z/x}^{\infty} f(x, y) dx dy + \int_0^{\infty} \int_{-\infty}^{z/x} f(x, y) dx dy.$$

En dérivant cette expression par rapport à z on a :

$$f^Z(z) = - \int_{-\infty}^0 \frac{1}{x} f(x, \frac{z}{x}) dx + \int_0^{\infty} \frac{1}{x} f(x, \frac{z}{x}) dx.$$

Le même raisonnement pour $\varphi(X, Y) = X + Y$ conduit à :

$$F^Z(z) = \int_{-\infty}^{\infty} \left\{ \int_{-\infty}^{z-x} f(x, y) dy \right\} dx,$$

$$f^Z(z) = \int f(x, z-x) dx = \int f(z-y, y) dy.$$

Nous verrons ultérieurement que dans le cas où $f(x, y) = h(x)g(y)$, $f^Z(z)$ est appelée produit de convolution de h et g .

3. Espérance, covariance

Etant donné un vecteur aléatoire X dans \mathbb{R}^d , l'espérance et la variance d'une composante X_i , si elles existent, sont définies par

$$E(X_i) = \int_{\mathbb{R}} x P^{X_i}(dx) = \int_{\mathbb{R}^d} x_i dF^X(x_1, \dots, x_d)$$

$$\text{Var}(X_i) = \int_{\mathbb{R}} (x - E(X_i))^2 P^{X_i}(dx) = \int_{\mathbb{R}^d} (x_i - E(X_i))^2 dF^X(x_1, \dots, x_d).$$

En particulier, si la loi du vecteur aléatoire X admet une densité $f(x_1, \dots, x_d)$ alors

$$E(X_i) = \int_{\mathbb{R}^d} x_i f(x_1, \dots, x_d) dx_1 \cdots dx_d,$$

$$\text{Var}(X_i) = \int_{\mathbb{R}^d} (x_i - E(X_i))^2 f(x_1, \dots, x_d) dx_1 \cdots dx_d.$$

On définit donc un vecteur moyenne $E(X)$ dont chaque composante est la moyenne de la composante correspondante du vecteur X . Pour tenir compte du lien éventuel entre les diverses composantes du vecteur aléatoire, nous sommes conduit à introduire la notion de covariance de deux variables aléatoires. Soit X et Y deux variables aléatoires réelles, alors la covariance $\text{Cov}(X, Y)$ du couple (X, Y) est définie par

$$\text{Cov}(X, Y) = E[(X - EX)(Y - EY)].$$

Supposons que X et Y soient de carré intégrable, alors d'après l'inégalité de Cauchy-Schwarz $|\text{Cov}(X, Y)| \leq \text{Var}(X)^{1/2} \text{Var}(Y)^{1/2}$, ce qui conduit à définir le coefficient de corrélation par

$$\rho_{XY} := \frac{\text{Cov}(X, Y)}{\text{Var}(X)^{1/2} \text{Var}(Y)^{1/2}}, \quad -1 \leq \rho_{XY} \leq 1.$$

Lorsque ce coefficient de corrélation est nulle, on dira que les variables aléatoires ne sont pas corrélées.

Définition 3.1. Soit $X = (X_1, \dots, X_d)$ un vecteur aléatoire sur un espace probabilisé (Ω, \mathcal{A}, P) . On dit que X est de puissance p -ième intégrable si chacune de ses composantes l'est. Son espérance est le vecteur de \mathbb{R}^d

$$E(X) = (E(X_1), \dots, E(X_d)).$$

Sa matrice (carrée) de covariance est

$$\text{Cov}(X) = \left(E((X_i - EX_i)(X_j - EX_j)) \right)_{1 \leq i, j \leq d}.$$

A la variance d'une variable aléatoire réelle se substitue une matrice carrée C symétrique et qui est semi-définie positive dans le sens suivant : pour tous les réels $\lambda_1, \dots, \lambda_d$,

$$\sum_{1 \leq i, j \leq d} \lambda_i \lambda_j C_{ij} \geq 0.$$

On effet, on vérifie facilement que dans le cas d'une matrice de covariance, le terme de droite est égal à

$$E \left[\left(\sum_{1 \leq i \leq d} \lambda_i (X_i - EX_i) \right)^2 \right].$$

Exemple 40 (Vecteur gaussien). Soit le vecteur aléatoire X à valeurs dans \mathbb{R}^d dont la densité par rapport à la mesure de Lebesgue sur \mathbb{R}^d est donnée par

$$f(x) = \frac{1}{(2\pi)^{d/2}} \exp(-\|x\|^2/2),$$

avec $x = (x_1, \dots, x_d)$ et la norme euclidienne $\|x\|^2 = x_1^2 + \dots + x_d^2$. D'après le théorème de Fubini, les lois marginales de chaque composante sont des lois $\mathcal{N}(0, 1)$. La moyenne de ce vecteur est le vecteur nul et la matrice de covariance est la matrice identité. On étudiera au chapitre suivant les vecteurs gaussiens de moyenne un vecteur m et de matrice de covariance une matrice symétrique définie positive C .

Exemple 41 (Loi de Cauchy symétrique dans \mathbb{R}^2). Considérons la fonction de deux variables

$$f(x_1, x_2) = \frac{1}{2\pi} \frac{1}{\sqrt{(1+x_1^2+x_2^2)^3}}.$$

Pour vérifier que cette fonction définit une densité sur \mathbb{R}^2 , notons que

$$\int_{-\infty}^{\infty} f(x_1, y) dy = \frac{1}{2\pi} \frac{1}{1+x_1^2} \frac{y}{\sqrt{1+x_1^2+y^2}} \Big|_{-\infty}^{\infty} = \frac{1}{\pi} \frac{1}{1+x_1^2}.$$

Par conséquent, la loi marginale de X_1 (resp. X_2) est une loi de Cauchy et donc f est bien une densité.

Effectuons un changement de variable en introduisant les v.a.r. R et Θ telles que

$$X_1 = R \cos \Theta, \quad X_2 = R \sin \Theta, \quad R \geq 0, \quad 0 \leq \Theta < 2\pi.$$

Pour calculer la densité du couple (R, Θ) , appliquons la formule du changement de variable, d'où

$$f^{(R, \Theta)}(r, \theta) = f(r \cos \theta, r \sin \theta)r.$$

Dans le cas de la loi de Cauchy symétrique, on obtient que le couple (X_1, X_2) est un vecteur de direction choisie uniformément sur $[0, 2\pi[$ et de longueur R de densité $r(1 + r^2)^{-3/2}$.

4. Fonction caractéristique

De la même façon que dans le cas des v.a.r., on peut définir la fonction caractéristique d'un vecteur aléatoire, celle-ci caractérisant la loi du vecteur.

Définition 4.1. Soit X un vecteur aléatoire sur (Ω, \mathcal{A}, P) à valeurs dans \mathbb{R}^d . On appelle fonction caractéristique de X ou de la loi de X , la fonction à valeurs complexes, notée φ^X

$$t \in \mathbb{R}^d \mapsto \varphi^X(t) = E(e^{i\langle t, X \rangle}) = \int_{\mathbb{R}^d} e^{i\langle t, x \rangle} P^X(dx),$$

où $\langle t, x \rangle$ désigne le produit scalaire euclidien $t_1x_1 + \dots + t_dx_d$.

La fonction caractéristique est à valeurs complexes, de module majoré par 1 (d'après l'inégalité de Jensen), et $\varphi^X(0) = 1$. Si la loi de X a une densité f par rapport à la mesure de Lebesgue sur \mathbb{R}^d , alors

$$\varphi^X(t) = \int_{\mathbb{R}^d} e^{i\langle t, x \rangle} f(x) dx,$$

est appelée transformée de Fourier de f . Notons également qu'on peut obtenir la fonction caractéristique de la composante X_i à partir de φ^X ; il suffit de choisir $t = (0, \dots, 0, t_i, 0, \dots, 0)$.

Comme dans le cas d'une v.a.r., la fonction caractérise la loi de X , dans le sens où si X et Y sont deux vecteurs aléatoires de loi P^X et P^Y telles que $\varphi^X = \varphi^Y$, alors $P^X = P^Y$.

Exemple 42. Si X est un vecteur aléatoire à valeurs dans \mathbb{R}^d , Σ une matrice opérant sur \mathbb{R}^d et $m \in \mathbb{R}^d$, alors $Y = \Sigma X + m$ est un vecteur aléatoire de fonction caractéristique

$$\varphi^Y(t) = e^{i\langle t, m \rangle} \varphi^X(\Sigma^T t),$$

puisque $\langle t, \Sigma X + m \rangle = \langle \Sigma^T t, X \rangle + \langle t, m \rangle$.

Enfin, si φ^X est indéfiniment dérivable, alors

$$\frac{\partial^{i_1 + \dots + i_d} \varphi^X}{\partial t_1^{i_1} \partial t_2^{i_2} \dots \partial t_d^{i_d}}(0) = i^{i_1 + \dots + i_d} \int x_1^{i_1} \dots x_d^{i_d} P^X(dx) = i^{i_1 + \dots + i_d} E(X_1^{i_1} \dots X_d^{i_d}).$$

5. Indépendance

Dans ce paragraphe, nous allons étendre la notion d'indépendance introduite au chapitre 1 pour une famille d'événements, à une famille de variables aléatoires réelles. Plus précisément, nous avons défini la notion d'indépendance d'une famille de sous-tribus de \mathcal{A} . Cette extension nous permet de définir aisément la notion d'indépendance d'une famille de v.a.r. $(X_i)_{i \in I}$ en considérant la famille de sous-tribus engendrées par cette famille de v.a.r., c.-à-d. par les $\mathcal{A}_i = \sigma(\{X_i^{-1}(A) : A \in \mathcal{B}(\mathbb{R})\})$.

Définition 5.1. *Une famille quelconque de variables aléatoires $(X_i)_{i \in I}$, définie sur (Ω, \mathcal{A}, P) et à valeurs dans (E, \mathcal{E}) , est (mutuellement) indépendante si la famille des tribus engendrées par les X_i est (mutuellement) indépendante, c.-à-d. pour tout $J \subset I$ fini, et tous les ensembles mesurables $B_j \in \mathcal{E}$, $j \in J$*

$$P\{X_j \in B_j : j \in J\} = P\left(\bigcap_{j \in J} \{X_j \in B_j\}\right) = \prod_{j \in J} P\{X_j \in B_j\}.$$

Exemple 43. Poursuivons l'exemple 1-(19). La famille de variables aléatoires

$$X_n = I_{A_n} = \sum_{1 \leq k \leq 2^{n-1}} I_{]2(k-1)2^{-n}, (2k-1)2^{-n}[}$$

de $[0, 1]$ dans $\{0, 1\}$ est indépendante. On vérifie facilement que la loi de X_n est une loi de Bernoulli de paramètre $1/2$.

Reformulons maintenant l'indépendance des variables aléatoires en terme de lois de ces variables. Rappelons que si $X = (X_1, \dots, X_d)$ est un vecteur aléatoire sur (Ω, \mathcal{A}, P) , la loi de X détermine la loi de chacune des variables X_i , mais que la réciproque est fautive en général. Néanmoins, si les coordonnées sont indépendantes, le résultat suivant montre que la loi du vecteur est déterminée par celle des marginales.

Proposition 5.2. *Soit (X_1, \dots, X_d) une famille finie de v.a.r. indépendantes sur (Ω, \mathcal{A}, P) . La loi P^X du vecteur aléatoire $X = (X_1, \dots, X_d)$ sur $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ est égale au produit des lois marginales $P^{X_1} \otimes \dots \otimes P^{X_d}$. Réciproquement, si la loi du vecteur est égale au produit des marges, alors les variables sont indépendantes.*

Ce résultat explique la terminologie introduite dans la remarque terminant l'exemple 36.

DÉMONSTRATION. Si $B = B_1 \times \dots \times B_d$ est un pavé dans $\mathcal{B}(\mathbb{R}^d)$, l'hypothèse d'indépendance donne,

$$\begin{aligned} P^X(B) &= P\{X_1^{-1}(B_1) \cap \dots \cap X_d^{-1}(B_d)\} \\ &= P\{X_1^{-1}(B_1)\} \dots P\{X_d^{-1}(B_d)\} \\ &= P^{X_1}(B_1) \dots P^{X_d}(B_d). \end{aligned}$$

L'identité s'étend à l'algèbre des réunions finies disjointes de pavés, laquelle engendre la tribu borélienne produit $\mathcal{B}(\mathbb{R}^d)$. La réciproque découle des

identités précédentes et de la définition d'une loi, puisque

$$\begin{aligned} P\{(X_1, \dots, X_d) \in B_1 \times \dots \times B_d\} &= P^X(B) \\ &= P^{X_1}(B_1) \dots P^{X_d}(B_d) \\ &= \prod_{1 \leq i \leq d} P\{X_i \in B_i\}. \end{aligned}$$

□

Exemple 44. Soit $(X, Y) \in \mathbb{R}^2$ un couple de variables aléatoires, de densité $f(x)g(y)$ par rapport à la mesure de Lebesgue sur \mathbb{R}^2 . Alors X et Y sont indépendantes et de densités respectives f et g si $\int f(x)dx = 1$.

On démontre le critère d'indépendance suivant (cf. [1]) :

Proposition 5.3. *Une famille quelconque de variables aléatoires réelles $X_i, i \in I$, sur (Ω, \mathcal{A}, P) , est indépendante si et seulement si pour toute partie finie $J \subset I$ et toute famille finie de fonctions boréliennes $\phi_i, i \in I$, telles que $\phi_i(X_i)$ soit intégrable pour tout $i \in I$,*

$$E\left(\prod_{i \in J} \phi_i(X_i)\right) = \prod_{i \in J} E(\phi_i(X_i)).$$

Il est souvent plus facile de vérifier ce critère, en calculant des espérances, que de vérifier directement l'indépendance d'une famille de v.a.r. à partir de la définition. Comme cas particulier de la proposition 5.3, notons que si X_1, \dots, X_n sont des variables indépendantes et intégrables,

$$E(X_1 \dots X_n) = E(X_1) \dots E(X_n).$$

Attention, cette propriété ne caractérise pas, en général, l'indépendance. Elle décrit en fait une propriété plus faible, dite de non corrélation.

Définition 5.4. *Deux v.a.r. $X, Y \in L^2(\Omega, \mathcal{A}, P)$ sont dites non corrélées si*

$$E(XY) = E(X)E(Y),$$

ou, de façon équivalente, si $E((X - EX)(Y - EY)) = 0$.

Exemple 45. Soit X une v.a.r. de loi normale $\mathcal{N}(0, 1)$. Alors X et $Y = X^2$ sont non corrélées. En effet

$$E(XY) = E(X^3) = 0 = E(X)E(Y).$$

Cependant, X et Y ne sont pas indépendantes; par exemple

$$P\{X \geq 1, Y \geq 1\} = P\{X \geq 1\} \neq P\{X \geq 1\}P\{Y \geq 1\},$$

puisque $P\{Y \geq 1\} < 1$.

Pour des variables aléatoires non corrélées, on vérifie facilement le résultat suivant :

Proposition 5.5. *Soient X_1, \dots, X_n des v.a.r. deux-à-deux non corrélées. Elles vérifient alors l'identité,*

$$\text{Var}\left(\sum_{1 \leq i \leq n} X_i\right) = \sum_{1 \leq i \leq n} \text{Var}(X_i).$$

Nous en déduisons l'inégalité, dite de Bienaymé-Tchebitchev

$$P\left\{\left|\sum_{1 \leq i \leq n} (X_i - EX_i)\right| \geq t\right\} \leq \frac{1}{t^2} \sum_{1 \leq i \leq n} \text{Var}(X_i), \quad t > 0.$$

DÉMONSTRATION. La première partie se démontre par un calcul trivial. L'inégalité de Bienaymé-Tchebitchev découle de l'inégalité de Tchebitchev (cf. exemple 2-(34) (ii)). \square

Nous déduisons également de la proposition 5.3, un critère d'indépendance utilisant les fonctions caractéristiques.

Corollaire 5.6. *La famille (X_1, \dots, X_n) de variables aléatoires réelles est indépendante si et seulement si pour tout $t_1, \dots, t_n \in \mathbb{R}^n$,*

$$\varphi^{(X_1, \dots, X_n)}(t_1, \dots, t_n) = \varphi^{X_1}(t_1) \cdots \varphi^{X_n}(t_n).$$

DÉMONSTRATION. Le produit $\varphi^{X_1}(t_1) \cdots \varphi^{X_n}(t_n)$ est la fonction caractéristique de la loi produit $P^{X_1} \otimes \cdots \otimes P^{X_n}$. Le résultat découle alors de la détermination de la loi par la fonction caractéristique. \square

6. Somme de variables aléatoires indépendantes

L'étude des sommes des variables aléatoires est fondamentale en théorie des probabilités, puisqu'elle conduit aux théorèmes limites, tels que la loi des grands nombres et le théorème central limite. Par exemple, d'après la proposition 5.5, si les v.a.r. sont indépendantes et de même loi, alors

$$P\left\{\left|\sum_{1 \leq i \leq n} (X_i - EX_i)\right| \geq t\sqrt{n}\right\} \leq \frac{\text{Var}(X_1)}{t^2}.$$

Ainsi, la somme $\sum_{1 \leq i \leq n} X_i$ est proche d'un terme déterministe, $\sum_{1 \leq i \leq n} EX_i = nEX_1$, plus un terme aléatoire de l'ordre au plus \sqrt{n} .

L'objet de ce paragraphe est d'évaluer la loi de la somme $\sum_{1 \leq i \leq n} X_i$, ce qui nous permettra ultérieurement de préciser le comportement du terme aléatoire de l'ordre de \sqrt{n} (théorème central limite). Notons également que pour tout $\epsilon > 0$ et tout n

$$P\left\{\left|\frac{1}{n} \sum_{1 \leq i \leq n} X_i - EX_1\right| \geq \epsilon\right\} \leq \frac{\text{Var}(X_1)}{n\epsilon^2},$$

et par conséquent, pour tout $\epsilon > 0$

$$\lim_{n \rightarrow \infty} P\left\{\left|\frac{1}{n} \sum_{1 \leq i \leq n} X_i - EX_1\right| \geq \epsilon\right\} = 0.$$

Ce résultat traduit la loi faible des grands nombres (faible puisque la convergence a lieu en probabilité), c.-à-d. qu'asymptotiquement EX_1 est égale à la moyenne empirique $n^{-1} \sum_{1 \leq i \leq n} X_i$. En particulier, si $X_i = I_A$, indicatrice d'un événement de \mathcal{A} , alors $P(A)$ est asymptotiquement égale à la fréquence des réalisations de l'événement A . Ceci est cohérent avec la notion intuitive des probabilités, montrant que le formalisme mathématique de la théorie des probabilités est raisonnable. Nous verrons plus en détail au chapitre 6, les différentes notions de convergence d'une suite de variables aléatoires et les lois des grands nombres.

Proposition 6.1. *Soient X et Y deux variables aléatoires réelles, indépendantes sur (Ω, \mathcal{A}, P) . La loi de la somme $X + Y$ est donnée par le produit de convolution $P^X * P^Y$ des lois P^X et P^Y , défini, pour toute fonction borélienne bornée ϕ de \mathbb{R} dans \mathbb{R} , par*

$$\begin{aligned} \int_{\mathbb{R}} \phi(x) P^X * P^Y(dx) &= \int_{\mathbb{R}} \left(\int_{\mathbb{R}} \phi(x+y) P^Y(dy) \right) P^X(dx) \\ &= \int_{\mathbb{R}} \left(\int_{\mathbb{R}} \phi(x+y) P^X(dx) \right) P^Y(dy). \end{aligned}$$

DÉMONSTRATION. On utilise le théorème de transport appliqué à la fonction $U(x, y) = x + y$ de $(\mathbb{R}^2, \mathcal{B}(\mathbb{R}^2), P^{(X,Y)})$ dans $(\mathbb{R}, \mathcal{B}(\mathbb{R}), P^{X+Y})$. Ainsi,

$$\begin{aligned} \int_{\mathbb{R}} \phi(z) P^{X+Y}(dz) &= E(\phi(X + Y)) \\ &= E(\phi(U(X, Y))) \\ &= \int_{\mathbb{R}^2} \phi \circ U(x, y) P^{(X,Y)}(dx, dy), \quad (\text{par transport}) \\ &= \int_{\mathbb{R}^2} \phi \circ U(x, y) (P^X \otimes P^Y)(dx, dy), \quad (\text{par indépendance}) \\ &= \int_{\mathbb{R}^2} \phi(x+y) P^X(dx) P^Y(dy). \end{aligned}$$

□

Remarque 6.2. L'opération de convolution des mesures vérifie un certain nombre de propriétés algébriques issues de la description en terme de variables aléatoires. En particulier :

- (i) $P * \delta_0 = P$ (puisque $X + 0 = X$),
- (ii) (commutativité) $P * Q = Q * P$,
- (iii) (associativité) $(P * Q) * R = P * (Q * R)$,
- (iv) (distributivité) $P * (\lambda Q + (1 - \lambda)R) = \lambda(P * Q) + (1 - \lambda)(P * R)$, pour tout $\lambda \in [0, 1]$.

Exemple 46. (i) Soient X et Y des variables aléatoires indépendantes à valeurs dans \mathbb{N} . Les lois respectives de X et Y s'expriment de la façon

suivante :

$$P^X = \sum_{n \geq 0} p_n^X \delta_n, \quad P^Y = \sum_{n \geq 0} p_n^Y \delta_n,$$

où $p_n^X := P(X = n)$ (idem pour p_n^Y) et δ_n désigne la mesure de Dirac en n . Notons tout d'abord que $\delta_n * \delta_m = \delta_{n+m}$. Par conséquent, en utilisant les propriétés de la convolution

$$\begin{aligned} P^X * P^Y &= \sum_{n \geq 0} \sum_{m \geq 0} p_n^X p_m^Y \delta_n * \delta_m \\ &= \sum_{n \geq 0} \left(\sum_{m=0}^n p_{n-m}^X p_m^Y \right) \delta_n \\ &= P^{X+Y}. \end{aligned}$$

Ainsi $P\{X + Y = n\} = \sum_{m=0}^n p_{n-m}^X p_m^Y$. Bien entendu, ce résultat peut être obtenu directement par un raisonnement probabiliste. Il suffit d'écrire que pour tout $n \in \mathbb{N}$, l'événement $\{X + Y = n\}$ est la réunion disjointe des événements $\{X + m = n, Y = m\}$, pour tout $0 \leq m \leq n$. Ainsi, par indépendance de X et Y ,

$$\begin{aligned} P\{X + Y = n\} &= \sum_{0 \leq m \leq n} P\{X + m = n, Y = m\} \\ &= \sum_{0 \leq m \leq n} P\{X = n - m\} P\{Y = m\} \\ &= \sum_{m=0}^n p_{n-m}^X p_m^Y. \end{aligned}$$

On peut vérifier à titre d'exemple que si X et Y suivent respectivement des lois de Poisson de paramètres λ et μ alors $X + Y$ suit une loi de Poisson de paramètre $\lambda + \mu$.

- (ii) Soient maintenant X et Y des v.a.r. admettant pour densités respectives, par rapport à la mesure de Lebesgue, les fonctions f et g . Alors, $X + Y$ admet une densité h par rapport à la mesure de Lebesgue, donnée par le produit de convolution des fonctions f et g , c.-à-d.

$$h(x) = f * g(x) = \int_{\mathbb{R}} f(x - y)g(y)dy = \int_{\mathbb{R}} g(x - y)f(y)dy, \quad x \in \mathbb{R}.$$

En effet, si ϕ est une fonction borélienne bornée,

$$\begin{aligned} \int \phi(x) P^X * P^Y(dx) &= \iint \phi(x + y) f(x) g(y) dx dy \\ &= \iint \phi(z) f(z - y) g(y) dz dy \\ &= \int \phi(z) h(z) dz. \end{aligned}$$

En particulier, on vérifie que $\mathcal{N}(0, \sigma_1^2) * \mathcal{N}(0, \sigma_2^2) = \mathcal{N}(0, \sigma_1^2 + \sigma_2^2)$.

- (iii) Nous avons toujours raisonner en termes de loi des v.a.r. Bien entendu, les résultats ci-dessus admettent une transcription en termes de fonctions de répartition. A savoir que la fonction de répartition d'une somme de deux v.a.r. X et Y indépendantes s'écrit comme le produit de convolution des fonctions de répartition de X et Y , c.-à-d.

$$F^{X+Y}(x) = \int_{-\infty}^{\infty} F^X(x-y)dF^Y(y) = \int_{-\infty}^{\infty} F^Y(x-y)dF^X(y),$$

où l'intégrale est prise au sens de Stieltjes. En particulier, si X admet une densité f alors $X + Y$ admet pour densité la fonction h obtenue par convolution entre f et la fonction de répartition F^Y , c.-à-d.

$$h(x) = \int_{-\infty}^{\infty} f(x-y)dF^Y(y).$$

Les fonctions caractéristiques fournissent un autre moyen, extrêmement puissant, de déterminer la loi de la somme de deux variables indépendantes.

Proposition 6.3. *Si X et Y sont deux variables aléatoires réelles indépendantes sur (Ω, \mathcal{A}, P) , la fonction caractéristique de leur somme est donnée par le produit des fonctions caractéristiques*

$$\varphi^{X+Y}(t) = \varphi^X(t)\varphi^Y(t), \quad t \in \mathbb{R}.$$

DÉMONSTRATION. Ce résultat est une conséquence de la proposition 6.1, puisque pour tout $t \in \mathbb{R}$,

$$\varphi^{X+Y}(t) = E(e^{it(X+Y)}) = E(e^{itX})E(e^{itY}) = \varphi^X(t)\varphi^Y(t).$$

□

On prendra soin de ne pas confondre la fonction caractéristique d'un couple (X, Y) de v.a.r. indépendantes, donnée par $\varphi^{(X,Y)}(s, t) = \varphi^X(s)\varphi^Y(t)$, $s, t \in \mathbb{R}$, avec la fonction caractéristique de la somme $X + Y$ qui ne dépend que d'une variable. Bien entendu, par définition de la fonction caractéristique, on a pour tout t , $\varphi^{X+Y}(t) = \varphi^{(X,Y)}(t, t)$.

6.1. Compléments. Nous allons dans cette partie présenter quelques types importants de lois sur \mathbb{R} , en rapport avec la somme de v.a.r. indépendantes. Aucune démonstration ne sera fournie, le lecteur pouvant se reporter par exemple à [7][§VI].

Lois infiniment divisibles. Nous avons vu que la fonction de répartition d'une somme de n v.a.r. (X_i) indépendantes et identiquement distribuées, est égale au n -ième produit de convolution de la fonction de répartition de X_1 , ou encore que la fonction caractéristique de la somme est égale à la puissance n -ième de la fonction caractéristique de X_1 . On peut alors se poser le problème inverse, à savoir sous qu'elles conditions, la fonction de répartition d'une v.a.r. peut s'écrire sous la forme de n produits de convolution d'une même fonction de répartition? En particulier, la loi gaussienne $\mathcal{N}(0, 1)$ peut s'écrire comme le produit de convolution de n loi gaussiennes $\mathcal{N}(0, n^{-1})$. Ceci conduit à la définition suivante :

Définition 6.4. Soit X une v.a.r. sur (Ω, \mathcal{A}, P) , de fonction de répartition F . On dit que F (ou la loi de X) est *infiniment divisible*, si pour tout n il existe une fonction de répartition F_n telle que

$$F = \underbrace{F_n * \cdots * F_n}_{n \text{ fois}} := F_n^{n*}.$$

En d'autres termes, F est *indéfiniment divisible* si et seulement si pour tout n , elle est la fonction de répartition de la somme $S_n = X_{1,n} + \cdots + X_{n,n}$ de n v.a.r. indépendantes et de même fonction de répartition F_n .

Il convient de remarquer que dans la décomposition de S_n , la loi commune des variables $X_{i,n}$, $1 \leq i \leq n$ dépend de n , d'où la présence d'un double indice.

Cette définition peut également s'énoncer en terme de fonction caractéristique, à savoir qu'une fonction de répartition F est infiniment divisible si et seulement si pour tout n , il existe une fonction caractéristique φ_n telle que $\varphi^X(t) = (\varphi_n(t))^n$.

La notion de loi infiniment divisible a été introduite par B. de Finetti (1929). La forme de la fonction caractéristique d'une telle loi, lorsque la variance est finie a été trouvée par A. Kolmogorov (1932) et dans le cas général par P. Lévy (1934).

Exemple 47. Nous avons déjà vu la stabilité des lois de Poisson et des lois gaussiennes pour la convolution. Ainsi, ces lois sont infiniment divisibles. Notons que la fonction caractéristique d'une loi de Poisson $\mathcal{P}(\lambda)$ étant

$$\varphi_\lambda(t) = \exp(\lambda(e^{it} - 1)),$$

on vérifie aisément la propriété d'infinie divisibilité, puisque pour tout n , $\varphi_\lambda = (\varphi_{\lambda/n})^n$. La loi de Cauchy, de fonction caractéristique $e^{-|t|}$ fournit un autre exemple de loi infiniment divisible.

Une manière relativement simple de construire une loi infiniment divisible est la suivante. Soit une fonction de répartition quelconque F et soit $\alpha > 0$. Définissons la fonction de répartition

$$G(x) = e^{-\alpha} \sum_{k=0}^{\infty} \frac{\alpha^k}{k!} F^{k*}(x),$$

avec la convention $F^{0*}(x) = 0$ pour tout $x < 0$ et $F^{0*}(x) = 1$ pour tout $x \geq 0$. Soit φ la fonction caractéristique de F , alors la fonction caractéristique de G est infiniment divisible, puisque

$$\begin{aligned} \int_{\mathbb{R}} e^{itx} dG(x) &= e^{-\alpha} \sum_{k=0}^{\infty} \frac{(\alpha\varphi(t))^k}{k!} \\ &= \exp \alpha(\varphi(t) - 1). \end{aligned}$$

Une interprétation de cette loi, appelée loi de Poisson composée, en termes de somme de v.a.r. est la suivante. Soit (X_n) une suite v.a.r. indépendantes et

de même fonction de répartition F . Soit N une v.a. discrète de loi de Poisson $\mathcal{P}(\alpha)$ indépendante de la suite (X_n) . Définissons une nouvelle v.a.r.

$$Y(\omega) = \sum_{k=1}^{N(\omega)} X_k(\omega),$$

avec la convention $\sum_{1 \leq k \leq 0} = 0$. Y s'exprime donc comme une somme de v.a.r. comprenant un nombre aléatoire de termes. Ce type de somme est très courant dans les applications pratiques, par exemple l'accumulation de chocs aléatoires dont l'intensité est également aléatoire, l'accumulation de dégats ou de contraintes sur une structure mécanique, etc.

On vérifie facilement, en utilisant l'indépendance de N vis-à-vis de la suite (X_n) , que Y admet bien pour fonction caractéristique, la fonction φ ci-dessus (exercice). Notons cependant, qu'il existe des lois infiniment divisibles qui ne sont pas du type précédent ; par exemple les lois gaussiennes et les lois de Cauchy.

Soit F une fonction de répartition infiniment divisible de fonction caractéristique φ . Alors, il existe une suite de fonctions caractéristiques (φ_n) telle que $\varphi = \lim_n \varphi_n^n$, d'où en passant à la limite $\varphi = \lim_n \varphi_n^n$. Or $|\varphi_n(t)| \leq 1$, d'où $\varphi_n^n(t) = [1 - (1 - \varphi_n(t))]^n \approx e^{n[\varphi_n(t)-1]}$. Par conséquent, pour que la limite existe, il faut que $n[\varphi_n(t) - 1] \rightarrow \psi(t)$, où $\psi(t)$ est une fonction, d'où $\varphi = e^\psi$. En fait, on démontre bien le résultat suivant (cf [7][§XVII-1]) :

Soit φ_n une suite de fonctions caractéristiques. Pour qu'il existe une limite continue

$$\varphi(t) = \lim_{n \rightarrow \infty} \varphi_n^n(t),$$

il est nécessaire et suffisant que

$$n[\varphi_n(t) - 1] \rightarrow \psi(t),$$

avec ψ continue. Dans ce cas, $\varphi = e^\psi$.

Notons que pour tout $t > 0$, $e^{tn[\varphi_n(t)-1]} \rightarrow e^{t\psi} = \varphi^t$. Le terme de gauche représente la fonction caractéristique d'un processus de Poisson composé, donc φ^t est également une fonction caractéristique. On en conclut que φ est infiniment divisible. Ainsi, chaque loi infiniment divisible peut se représenter comme la limite de lois de Poisson composée. On prouve ainsi la caractérisation suivante des lois infiniment divisibles :

La classe des lois infiniment divisibles coïncide avec la classe des lois limites de lois de Poisson composée.

Il est également possible de déterminer la forme canonique de la fonction ψ , (formule de P. Lévy ou de Khintchine, cf. [7][§XVII.2]).

Lois stables sur \mathbb{R} . Une classe particulière de lois infiniment divisibles est constituée des lois stables, qui jouent un rôle important dans l'étude des lois asymptotiques de la somme de v.a.r. indépendantes. On pourrait introduire ces lois à partir de la forme de la fonction ψ (cf. [7][§XVII.3]), il est cependant plus instructif, d'un point de vue probabiliste, de les introduire à partir de la somme de v.a.r. indépendantes (cf. [7][§VI.1]).

Tout d'abord, étant données deux variables aléatoires X et Y , on notera

$$X \stackrel{d}{=} Y$$

pour indiquer que X et Y ont même loi. Ainsi $X \stackrel{d}{=} aY + b$ signifie que X et Y diffèrent (en loi) par un paramètre translation et un paramètre de changement d'échelle (on dit également dans ce cas, que les lois de X et de Y sont de même type). Soient X, X_1, X_2, \dots une suite de v.a.r. indépendantes et de fonction de répartition commune F , et $S_n = X_1 + \dots + X_n$.

Définition 6.5. *La fonction de répartition F est dite stable (au sens large) si pour tout n il existe des constantes $c_n > 0$ et γ_n , telles que*

$$S_n \stackrel{d}{=} c_n X + \gamma_n,$$

et F n'est pas concentrée en un point (c.-à-d. que X n'est pas p.s. constante). F est dite stable au sens stricte si la relation précédente est valable avec $\gamma_n = 0$.

Exemple 48. (i) Nous savons que si X est une v.a.r. de loi $\mathcal{N}(0, 1)$, alors la stabilité de cette loi par convolution implique que S_n est de loi $\mathcal{N}(0, n)$; par conséquent la loi normale $\mathcal{N}(0, 1)$ est stable au sens stricte avec $c_n = n^{1/2}$.

(ii) La loi de Cauchy est une loi stable avec $c_n = n$.

L'importance des lois stables provient de l'extension du théorème central limite. Comme nous le verrons ultérieurement, ce théorème dit qu'étant donnée une suite (X_n) de v.a.r. indépendantes et identiquement distribuées telle que $EX_1 = 0$ et $\text{Var}(X_1) = 1$, alors la loi de $n^{-1/2}S_n$ tend vers la loi $\mathcal{N}(0, 1)$. On peut alors se poser plus généralement le problème suivant : étant donnée une suite X_1, X_2, \dots , de v.a.r. indépendantes et de même loi quelconque, quelles sont les distributions limites possibles de $a_n^{-1}(S_n - b_n)$, où $a_n > 0$ et b_n sont des constantes convenablement choisies ? La solution de ce problème fut obtenue par P. Lévy en 1937, par l'invention des lois stables. En effet, les seules lois limites sont les lois stables (cf. [7]).

Remarque 6.6. (i) Une loi stable est infiniment divisible, puisque la définition implique que pour tout x et tout n , $F(x) = F^{n*}(c_n x + \gamma_n)$.

(ii) Supposons que $E(|X|) < \infty$, alors $\gamma_n = (n - c_n)EX$. Il suffit de centrer X pour faire disparaître le coefficient γ_n . D'autre part, par la loi des grands nombres (cf. chap. 6), $n^{-1}S_n$ converge p.s. vers EX quand $n \rightarrow \infty$, d'où $\lim n^{-1}c_n = 0$. Notons que pour la loi de Cauchy, $c_n = n$; d'où la conclusion que pour cette loi, l'espérance est infinie.

(iii) Si $\text{var}(X) < \infty$, alors $\text{var}(S_n) = n\text{var}(X) = c_n^2\text{var}(X)$; donc $c_n = n^{1/2}$.

Le coefficient c_n a une forme bien précise, puisqu'on montre ([7]) que :

La constante de normalisation est de la forme $c_n = n^{1/\alpha}$ avec $0 < \alpha \leq 2$.

La constante α est appelée exposant caractéristique de F . On montre également que les seules lois limites sont des lois stables et que la seule loi

stable d'exposant caractéristique $\alpha = 2$ est la loi normale. On en déduit par conséquent, que si X est de variance finie la seule loi limite possible est la loi normale. D'autre part d'après la remarque 6.6 (ii), les lois stables d'exposant $\alpha \leq 1$ ne peuvent avoir une espérance finie.

Qu'en est-il du coefficient γ_n ? Notons que la somme S_{nm} est distribuée comme la somme de m v.a.r. indépendantes, chacune distribuée comme $c_n X + \gamma_n$. Ainsi

$$S_{nm} \stackrel{d}{=} c_n S_m + m\gamma_n \stackrel{d}{=} c_n c_m X + c_n \gamma_m + m\gamma_n.$$

Les indices n et m jouant un rôle symétrique, on en déduit la relation

$$(c_n - n)\gamma_m = (c_m - m)\gamma_n.$$

Hormis le cas $\alpha = 1$, on obtient $\gamma_n = b(c_n - n)$. Notons que d'après une remarque précédente, si $E|X| < \infty$, alors $b = -EX$. Pour $\alpha = 1$, on a $\gamma_{mn} = m\gamma_n + n\gamma_m$, d'où en divisant par mn , la relation $\gamma_n = \gamma_n \log n$. On montre alors le résultat suivant :

Si F est strictement stable d'exposant α , alors pour tout $s > 0$ et $t > 0$ on a $s^{1/\alpha} X_1 + t^{1/\alpha} X_2 \stackrel{d}{=} (s+t)^{1/\alpha} X$.

Exemple 49 (distribution de Holtsmark). Etant donnée une population d'étoiles, qu'elle est la composante de la force de gravitation engendrée par ce champ d'étoile en un point O choisi au hasard et selon une certaine direction x ? L'idée sous-jacente est que le champ d'étoile est approximativement composé d'agrégats aléatoires de points dont la masse est également aléatoire. Ceci peut se formaliser en termes de distributions de Poisson, mais le problème peut être résolu sans cela.

Considérons que la densité du nombre d'étoiles λ soit une variable, et introduisons la composante X_λ du champ de gravitation correspondant selon l'axe x . Intuitivement, la distribution aléatoire des étoiles implique que deux agrégats indépendants de densité respective s et t peuvent être combinés en un seul agrégat d'intensité $s+t$. En termes probabilistes, ceci se traduit par

$$(6.1) \quad X_s + X_t \stackrel{d}{=} X_{s+t}.$$

A un changement de densité de 1 à λ correspond un changement d'échelle de 1 à $\lambda^{1/3}$. La force de gravitation étant fonction de l'inverse du carré de la distance, nous en déduisons que X_t a la même loi que $t^{2/3} X_1$. Par conséquent, compte-tenu du résultat précédent et de l'identité (6.1), x_λ admet pour densité une loi stable de paramètre $\alpha = 3/2$. Pour d'autres exemples voir [7] duquel est tiré celui-ci.

Espérance conditionnelle

1. Introduction

Etant donné un espace de probabilité (Ω, \mathcal{A}, P) , nous avons introduit la notion de probabilité conditionnelle

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \quad \text{si } P(B) \neq 0,$$

et noté que $P(\cdot|B)$ était une probabilité sur (Ω, \mathcal{A}) .

Considérons maintenant la variable aléatoire $X = I_A$, où A est un événement quelconque, alors

$$P(A \cap B) = E(I_A I_B) = \int_B X dP.$$

Or $P(A|B)$ est également l'espérance de X pour la probabilité conditionnelle $P(\cdot|B)$, que nous noterons par $E(X|B)$, au même titre que $P(A)$ est l'espérance de X pour la probabilité P .

On obtient ainsi pour toute variable aléatoire étagée, la relation

$$(1.1) \quad E(X|B) = \frac{1}{P(B)} \int_B X dP,$$

et par passage à la limite on définit, pour toute variable aléatoire X , son espérance par rapport à la probabilité conditionnelle $P(\cdot|B)$ par l'identité (1.1).

Dans la précédente définition de l'espérance conditionnelle, nous avons toujours supposé que $P(B) \neq 0$. Cette hypothèse est trop restrictive, en particulier si on considère des événements de la forme $\{Y = y\}$, où Y est une variable aléatoire admettant une densité par rapport à la mesure de Lebesgue. En effet, dans ce cas $P(Y = y) = 0$ et la définition précédente ne s'applique plus. On souhaite donc étendre la définition de l'espérance conditionnelle à un cadre plus général. Pour cela, avant de donner une définition générale, nous considérerons deux cas particuliers.

1.1. Conditionnement par rapport à une variable aléatoire discrète.

Soit Y une variable aléatoire discrète prenant ses valeurs dans l'ensemble $E = \{y_1, y_2, \dots, y_n\}$. Soit $A \in \mathcal{A}$ un événement, alors

$$P(A|Y = y_i) = \frac{P(A \cap \{Y = y_i\})}{P(Y = y_i)},$$

pourvu que $P(Y = y_i) \neq 0$. Cette probabilité conditionnelle dépend de y_i ; c'est donc une fonction de la valeur prise par Y .

Construisons la variable aléatoire Z définie par

$$\begin{aligned} Z(\omega) &= P(A|Y = y_i), \text{ si } Y(\omega) = y_i \text{ et } P(Y = y_i) \neq 0, \\ &= \sum_{i \in I^*} P(A|Y(\omega) = y_i) I_{\{Y=y_i\}}(\omega), \\ &= \sum_{i \in I^*} \frac{1}{P(Y = y_i)} \left(\int_{\{Y=y_i\}} I_A dP \right) I_{Y=y_i} \end{aligned}$$

où I^* est l'ensemble des y_i tels que $P(Y = y_i) \neq 0$. On obtient bien une variable aléatoire qui est définie à un ensemble de probabilité nulle près.

Soit M un événement appartenant à la sous-tribu $\sigma(Y)$ engendrée par Y , c.-à-d. qu'il existe un sous-ensemble S de E telle que

$$M = \{\omega : Y(\omega) \in S\} = \bigcup_{y_j \in S} \{Y(\omega) = y_j\}.$$

On a l'importante identité suivante

$$\begin{aligned} P(M \cap A) &= \sum_{y_j \in S} P(A|Y = y_j) P(Y = y_j) \\ &= \int_M Z dP. \end{aligned}$$

De façon habituelle, on étend les résultats précédents à une variable aléatoire étagée, puis à toute variable aléatoire X , et on obtient une variable aléatoire notée $E(X|Y)$ et appelée *espérance conditionnelle de X sachant Y* , définie par

$$E(X|Y) = \sum_{i \in I^*} \frac{1}{P(Y = y_i)} \left(\int_{\{Y=y_i\}} X dP \right) I_{Y=y_i}.$$

Notons que cette variable aléatoire est définie à un ensemble de probabilité nulle près et est mesurable par rapport à la tribu $\sigma(Y)$ engendrée par Y .

De la même façon pour tout événement $M \in \sigma(Y)$, on a

$$E(X I_M) = \int_M X dP = \sum_{y_j \in S} \int_{\{Y=y_j\}} X dP,$$

d'autre part

$$\begin{aligned} E(I_M E(X|Y)) &= \sum_{i \in I^*} \frac{1}{P(Y = y_i)} \left(\int_{\{Y=y_i\}} X dP \right) P(I_M \{Y = y_i\}) \\ &= \sum_{y_j \in S} \frac{P(Y = y_j)}{P(Y = y_j)} \left(\int_{\{Y=y_j\}} X dP \right) P(I_M \{Y = y_j\}) \\ &= E(X I_M). \end{aligned}$$

Exemple 50 ([1]). Soit X une variable aléatoire sur (Ω, \mathcal{A}, P) de loi de Poisson $\mathcal{P}(\lambda)$. Soit $Y = 2[X/2]$, où $[\cdot]$ désigne la fonction partie entière. Y

est donc une variable discrète prenant ses valeurs dans $\{2n, n \in \mathbb{N}\}$. On a donc

$$\begin{aligned} E(X|Y = 2n) &= \frac{1}{P(Y = 2n)} \int_{\{Y=2n\}} X dP \\ &= \frac{1}{P(X = 2n) + P(X = 2n + 1)} \left(\int_{\{X=2n\}} X dP + \int_{\{X=2n+1\}} X dP \right) \\ &= \frac{1}{e^{-\lambda} \left(\frac{\lambda^{2n}}{(2n)!} + \frac{\lambda^{2n+1}}{(2n+1)!} \right)} (2nP(X = 2n) + (2n + 1)P(X = 2n + 1)) \\ &= \frac{2n + 1)(2n + \lambda)}{2n + 1 + \lambda}, \end{aligned}$$

d'où

$$\begin{aligned} E(X|Y) &= \sum_{n \geq 0} \frac{2n + 1)(2n + \lambda)}{2n + 1 + \lambda} Y_{\{Y=2n\}} \\ &= \frac{(Y + \lambda)(Y + 1)}{Y + 1 + \lambda}. \end{aligned}$$

Notons que $E(X|Y)$ s'exprime comme une fonction mesurable de la variable aléatoire Y . Or ceci est une conséquence du fait que $E(X|Y)$ est mesurable par rapport à la tribu engendrée par Y (cf. Lemme de Doob 3.1 ci-dessous). Nous verrons que cette propriété s'étend au cas général où Y n'est plus nécessairement discrète. Il peut être parfois judicieux de poser a priori que $E(X|Y) = h(Y)$ et de calculer h , ceci permettant de garder en mémoire la forme du résultat à obtenir.

1.2. Conditionnement par rapport à une variable aléatoire continue. Dans cette partie, nous traiterons d'une façon non rigoureuse du conditionnement par rapport à une variable aléatoire Y admettant une densité $f^Y(y)$. Les expressions obtenues seront identiques à celles obtenues d'une façon rigoureuse, mais nous espérons de cette manière faire comprendre les généralisations et résultats obtenus dans le cadre le plus général, sachant que ceux-ci ne seront démontrés.

Soient donc X et Y deux variables aléatoires de densité jointe $f^{(X,Y)}(x, y)$, et considérons la probabilité conditionnelle de l'événement $\{X \leq x\}$ sachant que $y < Y \leq y + h$. Nous avons

$$\begin{aligned} P(X \leq x | y < Y \leq y + h) &= \frac{P(X \leq x; y < Y \leq y + h)}{P(y < Y \leq y + h)} \\ &= \frac{\int_y^{y+h} du \int_{-\infty}^x f^{(X,Y)}(v, u) dv}{\int_y^{y+h} f^Y(u) du}, \end{aligned}$$

où $f^Y(u)$ est la densité de la variable aléatoire Y .

Divisons le numérateur et le dénominateur par h , et faisons tendre h vers zéro. Nous obtenons ainsi une fonction $U_y(x)$ de x définie par

$$U_y(x) = \frac{1}{f^Y(y)} \int_{-\infty}^x f^{(X,Y)}(v, y) dv,$$

pouvant s'interpréter comme la probabilité conditionnelle de l'événement $\{X \leq x\}$ sachant que $Y = y$. Notons aussi que pour tout x fixé, $U_y(x)$ est une fonction de répartition de densité

$$u_y(x) = \frac{1}{f^Y(y)} f^{(X,Y)}(x, y).$$

On appellera u_y la densité conditionnelle de X sachant $Y = y$, et l'espérance conditionnelle de X sachant $Y = y$ est définie par

$$\begin{aligned} E(X|Y = y) &= \int_{-\infty}^{\infty} x dP(X \leq x|Y = y) dx \\ &= \int_{-\infty}^{\infty} x u_y(x) dx = \frac{1}{f^Y(y)} \int_{-\infty}^{\infty} x f^{(X,Y)}(x, y) dx. \end{aligned}$$

On notera que l'on obtient bien une fonction de y , et que l'évènement $\{Y = y\}$ est de probabilité nulle en raison de l'existence d'une densité $f^Y(y)$. Néanmoins, il faudrait en toute généralité définir une variable aléatoire $E(X|Y)$ qui soit $\sigma(Y)$ mesurable. Dans ce cas, $E(X|Y) = h(Y)$ où h est une fonction $\sigma(Y)$ -mesurable et telle que sous $h(y) = E(X|Y = y)$. Ce sera l'objet de la section suivante.

2. Conditionnement général

Nous allons maintenant définir rigoureusement la notion d'espérance conditionnelle d'une variable aléatoire X par rapport à une sous-tribu quelconque, de façon à retrouver les résultats précédents lorsque la sous-tribu sera engendrée par une variable aléatoire Y discrète ou admettant une densité.

Pour cela énonçons le théorème principal qui définit l'espérance conditionnelle.

Théorème 2.1. *Soit (Ω, \mathcal{A}, P) un espace probabilisé, et soit \mathcal{B} une sous-tribu de \mathcal{A} . Etant donnée une variable aléatoire réelle X sur (Ω, \mathcal{A}, P) et **intégrable**, alors il existe une unique (au sens p.s.) variable aléatoire, appelée espérance conditionnelle de X sachant \mathcal{B} et notée $E(X|\mathcal{B})$, telle que*

- (i) $\omega \mapsto E(X|\mathcal{B})(\omega)$ est \mathcal{B} -mesurable,
- (ii) pour tout $B \in \mathcal{B}$, $\int_B E(X|\mathcal{B}) dP = \int_B X dP$.

Remarque 2.2. Supposons que $X = I_A$ pour $A \in \mathcal{A}$, alors la propriété (ii) devient pour tout $B \in \mathcal{B}$

$$\int_B E(X|\mathcal{B}) dP = P(A \cap B).$$

On obtient ainsi une extension de la notion de probabilité conditionnelle par rapport à une sous-tribu.

En particulier, étant donnée une partition $(B_i)_{i \in I}$, $I \subset \mathbb{N}$ d'événements de \mathcal{A} , la tribu \mathcal{B} engendrée par cette partition est la collection de toutes les unions possibles d'événements B_i et de leurs complémentaires. Ainsi, tout ensemble $A \in \mathcal{B}$ peut être fractionné sur les ensembles élémentaires B_i .

On peut alors vérifier que la variable aléatoire

$$Z = \sum_{i \in I^*} \frac{1}{P(B_i)} \left(\int_{B_i} X dP \right) I_{B_i},$$

vérifie les conditions du théorème 2.1 d'où, en raison de l'unicité, Z est une représentation de l'espérance conditionnelle $E(X|\mathcal{B})$.

Si $X = I_A$, on note $E(X|\mathcal{B}) = P(A|\mathcal{B})$ et l'espérance conditionnelle est appelée probabilité conditionnelle de A par rapport à la sous-tribu \mathcal{B} , qui est une extension de la probabilité conditionnelle classique.

Si Y est une variable aléatoire discrète, et \mathcal{B} est la sous-tribu engendrée par Y , donc engendrée par les $B_i = \{y_i\}$, où les y_i sont les valeurs possibles de Y , on retrouve la version de l'espérance conditionnelle obtenue en introduction.

Il existe deux preuves de ce théorème. La première s'appuie sur les propriétés des projections orthogonales dans l'espace de Hilbert $L^2(\Omega, \mathcal{A}, P)$ [1], la seconde utilise le théorème de Radon-Nikodym [11]. Cette dernière est plus facile, la difficulté étant reportée sur la preuve du théorème de Radon-Nikodym. En effet (cf 2-(1.2.1)), si X est positive, la mesure $\nu(A) = \int_A X dP$ est absolument continue par rapport à P et donc sa restriction $\nu_{\mathcal{B}}$ à \mathcal{B} est également absolument continue par rapport à la restriction $P_{\mathcal{B}}$ de P à la même sous-tribu. Par conséquent, il existe une densité $E(X|\mathcal{B}) = d\nu_{\mathcal{B}}/dP_{\mathcal{B}}$. On obtient directement les deux propriétés du théorème. Pour X quelconque, on utilise la décomposition habituelle $X = X^+ - X^-$.

Un certain nombre de propriétés découlent immédiatement du théorème 2.1.

Propriétés :

- (i) $E(aX + bY + c|\mathcal{B}) = aE(X|\mathcal{B}) + bE(Y|\mathcal{B}) + c$ p.s.,
- (ii) si $X \leq Y$, alors $E(X|\mathcal{B}) \leq E(Y|\mathcal{B})$ p.s.,
- (iii) si X_n converge p.s. vers X en croissant, alors $E(X_n|\mathcal{B})$ converge p.s. en croissant vers $E(X|\mathcal{B})$,
- (iv) si $\varphi : \mathbb{R} \rightarrow \mathbb{R}$ est convexe et $\varphi(X)$ intégrable, alors $\varphi(E(X|\mathcal{B})) \leq E(\varphi(X)|\mathcal{B})$ p.s. (inégalité de Jensen). En particulier, $|E(X|\mathcal{B})| \leq E(|X||\mathcal{B})$ et $(E(X|\mathcal{B}))^2 \leq E(X^2|\mathcal{B})$,
- (v) si $\mathcal{B} = \{\Omega, \emptyset\}$, $E(X|\mathcal{B}) = E(X)$,
- (vi) si $\mathcal{C} \subset \mathcal{B} \subset \mathcal{A}$, $E(E(X|\mathcal{B})|\mathcal{C}) = E(X|\mathcal{C})$. Le conditionnement successif sera noté par la suite $E(X|\mathcal{B}|\mathcal{C})$,
- (vii) $E(E(X|\mathcal{B})) = E(X)$,
- (viii) si \mathcal{B} est indépendante de la sous-tribu $\sigma(X)$ engendrée par la v.a. X , alors $E(X|\mathcal{B}) = E(X)$ p.s.,
- (ix) si Y est \mathcal{B} -mesurable et XY est intégrable alors $E(XY|\mathcal{B}) = YE(X|\mathcal{B})$.

Nous n'explicitons pas la preuve des propriétés, exceptée pour les 3 dernières (cf. [1]).

DÉMONSTRATION. (vii) : il suffit de prendre $B = \Omega$ dans 2.1-(ii)
(viii) : si $B \in \mathcal{B}$, I_B et X sont indépendantes et donc pour tout $B \in \mathcal{B}$,

$$\int_B E(X|\mathcal{B})dP = \int I_B X dP = E(X)P(B).$$

La relation précédente implique que $\int_B [E(X|\mathcal{B}) - E(X)]dP = 0$ pour tout $B \in \mathcal{B}$. Puisque $E(X|\mathcal{B})$ est \mathcal{B} -mesurable, on obtient que $E(X|\mathcal{B}) = E(X)$ p.s.

(ix) Le résultat est clair si $Y = I_A$, $A \in \mathcal{B}$. En effet, nous avons

$$\begin{aligned} \int_B E(XY|\mathcal{B})dP &= \int_B XY dP = \int_{B \cap A} X dP \\ &= \int_{B \cap A} E(X|\mathcal{B})dP \\ &= \int_B Y E(X|\mathcal{B})dP. \end{aligned}$$

Nous avons utilisé dans la dernière ligne, le fait que $B \cap A \in \mathcal{B}$. Par un argument identique à la preuve de (viii), on en déduit que $E(XI_A|\mathcal{B}) = I_A E(X|\mathcal{B})$, pour tout $A \in \mathcal{B}$. Il suffit ensuite de suivre le schéma général de l'intégration, en approximant les v.a.r. positives par des fonctions étagées, puis en décomposant en partie positive et partie négative. \square

Notation : Si $\mathcal{B} = \sigma(Y)$ est la tribu engendrée par Y , on note $E(X|Y) = E(X|\mathcal{B})$. Si $X = (X_1, \dots, X_d)$ est un vecteur aléatoire, $E(X|\mathcal{B})$ est le vecteur $(E(X_1|\mathcal{B}), \dots, E(X_d|\mathcal{B}))$. Enfin si $X = I_A$, on note $P(A|\mathcal{B}) = E(I_A|\mathcal{B})$ qui est donc une extension de la probabilité conditionnelle par rapport à une sous-tribu quelconque.

3. Lois conditionnelles

Nous avons vu que dans le cas d'un couple (X, Y) de densité jointe $f^{(X,Y)}$, on avait une notion intuitive de l'espérance conditionnelle

$$E(X|Y = y) = \int x u_y(x) dx,$$

s'exprimant comme une espérance selon une densité de probabilité $u_y(x)$. D'autre part, le théorème 2.1 affirme l'existence d'une variable aléatoire $E(X|Y)$ mesurable par rapport à la tribu engendrée par Y . Nous allons dans la suite exhiber une fonction $h(Y)$ telle que $h(y) = E(X|Y = y)$ et une densité $K^Y(x)$ telle que

$$K^{Y=y}(x) = u_y(x) \quad \text{et} \quad h(y) = \int x K^{Y=y}(x) dx.$$

Pour cela énonçons, sans démonstration, le lemme suivant :

Lemme 3.1 (de Doob). *Soit Y une variable aléatoire sur (Ω, \mathcal{A}) , à valeurs réelles, et soit $X : \Omega \rightarrow \mathbb{R}$. Pour que X soit mesurable par rapport à la tribu $\sigma(Y)$, il faut et il suffit qu'il existe une application borélienne $h : \mathbb{R} \rightarrow \mathbb{R}$, telle que $X = h(Y)$.*

Considérons à présent un couple (X, Y) de v.a.r. sur (Ω, \mathcal{A}, P) , tel que X soit intégrable. L'espérance conditionnelle $E(X|Y)$ est $\sigma(Y)$ -mesurable, ainsi d'après le lemme de Doob, il existe une fonction borélienne h telle que $E(X|Y) = h(Y)$. On conviendra d'appeler $h(y)$, $y \in \mathbb{R}$, l'espérance conditionnelle de X sachant $Y = y$, que nous noterons par $h(y) := E(X|Y = y)$. Cette notation, intuitivement explicite, est abusive car en toute rigueur h n'est pas définie par $E(X|Y = y)$, puisque $P\{Y = y\}$ peut être nulle. De même définir $E(X|Y = y)$ par $h(y)$ n'est pas possible, puisque h n'est définie que P^Y presque partout et en général $P\{Y = y\}$ est nul. La fonction h est seulement l'unique classe d'équivalence (à l'égalité P^Y p.s. près) de fonctions boréliennes telle que :

$$\forall B \in \mathcal{B}(\mathbb{R}), \int_{Y \in B} X dP = \int_B h(y) dP^Y(y).$$

3.1. Conditionnement par rapport à une v.a. discrète. Supposons que Y prenne un nombre fini ou dénombrable de valeurs y_i , $i \in I$, alors l'espérance conditionnelle est

$$E(X|Y) = \sum_{i \in I^*} \frac{1}{P(Y = y_i)} \left(\int_{Y=y_i} X dP \right) I_{Y=y_i},$$

où $I^* = \{i \in I : P\{Y = y_i\} > 0\}$. Nous avons alors pour tout $i \in I^*$,

$$h(y_i)E(X|Y = y_i) = \frac{1}{P(Y = y_i)} \int_{Y=y_i} X dP,$$

cette fonction n'étant pas définie sur $I \setminus I^*$.

3.2. Cas d'un couple de v.a.r. admettant des densités. Supposons que la loi du couple $(X, Y) \in \mathbb{R}^2$ admette une densité $f(x, y)$ par rapport à la mesure de Lebesgue. On va montrer que l'on peut choisir, conformément au calcul effectué en introduction de ce paragraphe,

$$(3.1) \quad h(y) = E(X|Y = y) = \frac{\int_{\mathbb{R}} x f(x, y) dx}{\int_{\mathbb{R}} f(x, y) dx},$$

lorsque $\int f(x, y) dx \neq 0$. Soit B un borélien et $A = Y^{-1}(B)$, donc $A \in \sigma(Y)$. Puisque Y admet pour densité $f^Y(y) = \int_{\mathbb{R}} f(x, y) dx$, on obtient

$$\begin{aligned} \int_A h(Y) dP &= \int_{y \in B} h(y) f^Y(y) dy \\ &= \int_{\mathbb{R}} \int_{y \in B} x f(x, y) dx dy \\ &= E(XI_A). \end{aligned}$$

Donc $h(Y)$ vérifie les propriétés (i) et (ii) du théorème 2.1, et par unicité p.s. nous avons $h(Y) = E(X|Y)$.

En remplaçant X par $\phi(X)$, où ϕ est une fonction borélienne bornée, on obtient

$$(3.2) \quad E(\phi(X)|Y) = \frac{\int \phi(x)f(x, Y)dx}{\int f(x, Y)dx} := \int \phi(x)K^Y(x)dx,$$

où $K^Y(x) = f(x, Y)/f^Y(Y)$. Choisissons $\phi(X) = I_B(X)$, avec $B \in \mathcal{B}(\mathbb{R})$, alors l'équation 3.2 donne

$$P(\{X \in B\}|Y) = \int_B K^Y(x)dx,$$

d'où l'interprétation de $K^Y(x)$ comme la densité par rapport à la mesure de Lebesgue de la probabilité conditionnelle $P(\cdot|Y)$. Il s'ensuit que la densité conditionnelle de X sachant $Y = y$ est $f(x, y)/f^Y(y)$, où f^Y est la densité de Y .

Définition 3.2. Soient X et Y deux v.a.r. définies sur (Ω, \mathcal{A}, P) . On appelle loi conditionnelle de X par rapport à Y sachant que $Y = y$, la probabilité (si elle existe) $K^y(dx)$ définie sur $\mathcal{B}(\mathbb{R})$ par

$$B \rightarrow P(\{X \in B\}|Y = y).$$

Si cette probabilité admet une densité $K^y(x)$, on dit que $K^y(x)$ est la densité conditionnelle de X par rapport à Y sachant $Y = y$.

Proposition 3.3. Soient (X, Y) un couple de v.a.r. admettant une densité jointe $f(x, y)$. Alors, la densité conditionnelle de X par rapport à Y sachant $Y = y$ existe et s'exprime par

$$K^Y(x)dx = \frac{f(x, Y)}{\int f(x, Y)dx}dx.$$

On note également la densité conditionnelle par $f(x|Y = y)$. En outre, pour toute fonction borélienne ϕ

$$E(\phi(X)) = \int f^Y(y)dy \int \phi(x)K^{Y=y}(x)dx.$$

La dernière assertion se prouve de la façon suivante :

$$\begin{aligned} E(\phi(X)) &= E(E(\phi(X)|Y)) \\ &= E\left(\int \phi(x)K^Y(x)dx\right) \\ &= \int f^Y(y)dy \int \phi(x)K^{Y=y}(x)dx. \end{aligned}$$

Cette dernière expression permet souvent de calculer l'espérance de $\phi(X)$ lorsque l'on connaît la loi de X sachant Y et la loi de Y , sans connaître la loi du couple (X, Y) . Plus généralement, il est possible d'obtenir la loi du

couple (X, Y) à partir de K^Y et de la loi de Y : si ϕ et ψ sont deux fonctions boréliennes bornées, alors

$$\begin{aligned} E(\phi(X)\psi(Y)) &= E(E(\phi(X)\psi(Y)|Y)) \\ &= E(\psi(Y)E(\phi(X)|Y)) \\ &= E(\psi(Y) \int \phi(x)K^Y(dx)). \end{aligned}$$

Remarque 3.4. La densité conditionnelle de X par rapport à Y sachant $Y = y$ s'exprime donc sous la forme $f(x, y) = f(x|Y = y)f^Y(y)$, qui n'est autre qu'une décomposition de la densité jointe du couple (X, Y) en un produit faisant intervenir la densité de la variable Y . Or, nous avons vu auparavant, que si X et Y sont indépendantes alors $f(x, y) = f^X(x)f^Y(y)$, d'où dans ce cas $f(x|Y = y) = f^X(x)$ (cf. propriété (viii)). En quelque sorte la décomposition de $f(x, y)$ à l'aide de la densité conditionnelle tient compte de la dépendance entre X et Y .

Notons également l'égalité utile

$$f(x, y) = f(x|Y = y)f^Y(y) = f(y|X = x)f^X(x),$$

qui permet connaissant la densité conditionnelle de X par rapport à Y sachant $Y = y$ d'en déduire la densité conditionnelle de Y par rapport à X sachant $X = x$ (formule de Bayes).

Exemple 51 ([10]). Soit (X, Y) un couple de v.a.r. tel que la loi conditionnelle de X sachant $Y = y$ est une loi de Poisson de paramètre y et que la loi de la variable Y admette pour densité :

$$f^Y(y) = ce^{-cy}, \quad y \geq 0, \quad c > 0.$$

Calculer $P(X = k)$ pour tout $k \in \mathbb{N}$ et la loi conditionnelle de Y sachant $X = k$.

D'après les relations précédentes, nous avons :

$$\begin{aligned} P(X = k) &= E(I_{\{X=k\}}) = \int_0^\infty E(I_{\{X=k\}}|Y = y)f^Y(y)dy \\ &= \int_0^\infty \frac{y^k}{k!}e^{-y}ce^{-cy}dy = \frac{c}{(1+c)^{k+1}}. \end{aligned}$$

Avant de calculer la loi conditionnelle de Y sachant X , calculons la loi du couple (X, Y) . On a pour toutes fonctions boréliennes bornées

$$E(\phi(X)\psi(Y)) = \int_0^\infty \sum_{k=0}^\infty \phi(k)\psi(y)\frac{y^k}{k!}ce^{-(1+c)y}dy,$$

d'où la loi jointe est donnée pour tout k par

$$P^{X,Y}(k, dy) = \frac{y^k}{k!}ce^{-(1+c)y}dy.$$

On obtient alors la loi conditionnelle $f(y|X = k)$ recherchée par

$$\begin{aligned} E(Y|X = k) &= \frac{1}{P(X = k)} \int I_{\{X=k\}}(x)y dP^{X,Y} \\ &= \frac{(1+c)^{1+k}}{c} \int_0^\infty y \frac{y^k}{k!} c e^{-(1+c)y} dy \\ &= \int y f(y|X = k) dy, \end{aligned}$$

d'où

$$f(y|X = k) = \frac{y^k}{k!} (1+c)^{1+k} e^{-(1+c)y}.$$

On aurait pu également écrire que la densité du couple (X, Y) est

$$f^{(X,Y)}(k, y) = P(X = k)f(y|X = k) = f^Y(y)f(k|Y = y)$$

d'où

$$\begin{aligned} f(y|X = k) &= \frac{f^Y(y)f(k|Y = y)}{P(X = k)} \\ &= \frac{c e^{-cy}}{c/(1+c)^{k+1}} \frac{y^k}{k!} e^{-y} \\ &= \frac{y^k}{k!} (1+c)^{1+k} e^{-(1+c)y}. \end{aligned}$$

3.3. Extension. On peut montrer un résultat plus général, qui permet d'étendre les expressions ci-dessus dans le cas de v.a.r. n'admettant pas nécessairement de densité. Bien que n'étant pas utilisé dans la suite du cours, ce résultat est fondamental et intervient en particulier dans l'étude des processus stochastiques et plus particulièrement les processus markoviens.

Définition 3.5. On appelle noyau de transition toute fonction $K : \mathbb{R} \times \mathcal{B}(\mathbb{R}) \rightarrow [0, 1]$ telle que

- (i) pour tout $B \in \mathcal{B}(\mathbb{R})$, la fonction $y \mapsto K(y, B) := K^y(B)$ est mesurable,
- (ii) pour tout $y \in \mathbb{R}$, la fonction d'ensembles $B \mapsto K^y(B)$ est une mesure de probabilité.

Le résultat suivant, dont une esquisse de la preuve figure dans [1], fournit l'existence d'un noyau de transition d'un couple de v.a.r. et généralise les exemples précédents :

Théorème 3.6. Soit (X, Y) un vecteur aléatoire dans $(\mathbb{R}^2, \mathcal{B}(\mathbb{R}^2))$ de loi P . Sous certaines hypothèses sur la tribu $\mathcal{B}(\mathbb{R}^2)$ (complétude par rapport à P), toujours supposées remplies dans les applications pratiques, il existe un noyau de transition K tel que pour toute fonction borélienne bornée ϕ

$$E(\phi(X)|Y) = \int \phi(x)K(\cdot, dx).$$

La mesure $K(y, dx)$ est appelée la loi conditionnelle de X sachant $Y = y$.

Ce résultat répond à notre problème du choix d'une version (régulière) de $E(X|Y)$. En effet, soit la fonction borélienne

$$\varphi(y) = \int \phi(x)K(y, dx),$$

qui existe d'après la définition du noyau de transition, alors on a

$$E(\phi(X)|Y = y) = \varphi(y).$$

Si la mesure $K(y, dx)$ admet une densité par rapport à la mesure de Lebesgue, celle-ci est appelée densité conditionnelle de X sachant $Y = y$, elle se note également $f(x|Y = y)$ et est une fonction en x, y .

Exemple 52. Soit X une v.a.r. et $h : \mathbb{R} \rightarrow \mathbb{R}$ mesurable ; quelle est la loi conditionnelle de $h(X)$ sachant $X = x$? Pour tout fonction borélienne bornée ϕ , nous avons $E(\phi(h(X))|X) = \phi(h(X))$. Par conséquent

$$\phi(h(x)) = \int \phi(y)K(x, dy) = \int \phi(y)\delta_{h(x)}(dy),$$

donc la loi de $h(X)$ sachant que $X = x$ est la masse de Dirac en $h(x)$.

4. Espérance conditionnelle et projection dans $L^2(\Omega, \mathcal{A}, P)$

Le but de ce paragraphe est de présenter une caractérisation géométrique de la notion d'espérance conditionnelle, caractérisation utile en particulier en théorie du filtrage.

Rappelons que si \mathcal{H}_0 est un sous-espace vectoriel fermé d'un espace de Hilbert \mathcal{H} , on peut projeter de façon unique tout élément de \mathcal{H} sur \mathcal{H}_0 de telle façon que :

$$\forall y \in \mathcal{H}, \exists z \text{ unique tel que } \forall x \in \mathcal{H}_0, y - z \perp x,$$

c.-à-d. que $\langle y - z, x \rangle = 0$, où $\langle \cdot, \cdot \rangle$ est le produit scalaire sur \mathcal{H} . Ce résultat n'est qu'une généralisation aux espaces de Hilbert de la notion de projection perpendiculaire dans \mathbb{R}^3 . Une autre caractérisation de la projection orthogonale est que z est aussi l'élément de \mathcal{H}_0 qui minimise la distance de z à y

$$\inf_{x \in \mathcal{H}_0} \|y - x\| = \|y - z\|.$$

Nous avons alors le résultat suivant :

Proposition 4.1. *Soit Y une v.a.r. de carré intégrable et soit \mathcal{B} une sous-tribu de \mathcal{A} . Soit \mathcal{H}_0 l'espace de Hilbert des classes d'équivalences de variables aléatoires de carré intégrable \mathcal{B} -mesurables. Alors, \mathcal{H}_0 est un sous-espace vectoriel fermé de $L^2(\Omega, \mathcal{A}, P)$ muni du produit scalaire hilbertien*

$$\langle X, Y \rangle = E(XY).$$

D'autre part $E(Y|\mathcal{B})$ est la projection orthogonale de Y sur \mathcal{H}_0 .

Le dernier résultat n'est qu'une reformulation de la propriété (ix) de l'espérance conditionnelle.

Si la sous-tribu \mathcal{B} correspond aux événements observés, alors une estimation \hat{Y} \mathcal{B} -mesurable de Y , minimisant la variance entre \hat{Y} et Y est précisément l'espérance conditionnelle. Comme application de cette interprétation, considérons le problème du filtrage. Supposons que nous observons un phénomène aléatoire dépendant du temps $X(t)$ (par ex. un bruit, les cours de la bourse, etc.) Soit $Z(s)$ les observations obtenues pour tout $s \leq t$, où t est fixé. En raison des incertitudes sur les mesures, nous avons la relation

$$Z(s) = X(s) + \text{bruit} .$$

Le problème du filtrage consiste à trouver le meilleur estimateur de $X(t)$ à partir des observations $Z(s)$ pour tout $s \leq t$. Intuitivement le problème consiste à filtrer le bruit pour le retirer d'une façon optimale des observations. Précisons ce que l'on entend par estimateur optimal. On dira que $\hat{X}(t)$ est un estimateur basé sur les observations $\{Z(s); s \leq t\}$ si $\hat{X}(t)$ est \mathcal{B}_t -mesurable, où \mathcal{B}_t est la sous-tribu engendrée par $\{Z(s); s \leq t\}$, c.-à-d. par tous les événements de la forme $Z^{-1}(s)(A)$ où A est un borélien de \mathbb{R} et ceci pour tout $s \leq t$. Cet estimateur sera dit optimal si

$$E(|X(t) - \hat{X}(t)|^2) = \inf\{E(|X(t) - Y|^2); Y \in \mathcal{K}\},$$

où \mathcal{K} désigne l'ensemble des v.a.r. de carré intégrable \mathcal{B}_t -mesurables. D'après le résultat ci-dessus, l'estimateur optimal $\hat{X}(t)$ est donc $E(X(t)|\mathcal{B}_t)$. Ceci est le point de départ de la théorie du filtrage.

Vecteurs aléatoires gaussiens

Pourquoi un chapitre particulier pour la loi gaussienne? Tout simplement, en raison de la quasi-universalité de la loi gaussienne qui s'est imposée dans de nombreuses disciplines (physique, génétique, biologie, etc.). La loi gaussienne (ou loi normale) est associée aux noms de Carl Friedrich Gauss et Pierre Simon de Laplace. La loi normale apparaît dans les travaux de Gauss sur l'estimation d'un paramètre à partir de diverses observations (méthodes des moindres carrés et du maximum de vraisemblance), tandis que Laplace l'obtient comme limite de la loi binomiale, ce qui constituera le point de départ de développements plus profonds en rapport avec la théorie des erreurs d'observations. L'aboutissement de ces travaux est le fameux théorème central limite. Pour plus de détails, se reporter à l'article de A. Fuchs dans [13].

1. La loi normale

Rappelons qu'une v.a.r. suit une loi normale ou loi gaussienne, $\mathcal{N}(m, \sigma^2)$, de moyenne m et de variance σ^2 , si sa densité par rapport à la mesure de Lebesgue sur \mathbb{R} est

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

La transformée de Fourier de cette densité est

$$\varphi^X(t) = \exp\left(imt - \frac{\sigma^2 t^2}{2}\right).$$

Rappelons également que si Y suit une loi $\mathcal{N}(0, 1)$, alors $X = m + \sigma Y$ suit une loi $\mathcal{N}(m, \sigma^2)$. Les paramètres m et σ^2 caractérisent complètement la loi normale. De plus cette loi admet des moments de tous ordres. Soit $\mu_p := E[(X - EX)]^p$, on obtient la formule de récurrence

$$\mu_p = (p-1)\sigma^2\mu_{p-2}, \quad p \geq 2.$$

Comme $\mu_1 = 0$, les moments d'ordre impair sont tous nuls, tandis que pour p pair

$$\mu_{2k} = (2k-1)!!\sigma^{2k} = \frac{(2k)!}{k!} \left(\frac{\sigma^2}{2}\right)^k,$$

où le symbole $(2k-1)!!$ désigne le produit de tous les nombres impairs de 1 à $2k-1$.

La fonction de répartition de la loi normale standard $\mathcal{N}(0, 1)$ est souvent notée

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-u^2/2} du.$$

Les valeurs de cette fonction sont tabulées dans des tables. Il est alors possible d'obtenir la probabilité qu'une v.a.r. X , de loi $\mathcal{N}(m, \sigma^2)$, soit dans l'intervalle $]a, b]$

$$P\{a < X \leq b\} = \Phi\left(\frac{b-m}{\sigma}\right) - \Phi\left(\frac{a-m}{\sigma}\right).$$

Explicitons quelques propriétés de la fonction Φ . Tout d'abord, en raison de la symétrie de la densité gaussienne, nous avons la relation

$$\Phi(-x) = 1 - \Phi(x),$$

d'où

$$P\{|X - m| \leq a\} = 2\Phi\left(\frac{a}{\sigma}\right) - 1.$$

En particulier, $\Phi(2) = 0,9750$ et $\Phi(3) = 0,9986$, d'où

$$P\{|X - m| > 2\sigma\} = 0,05, \quad P\{|X - m| > 3\sigma\} \approx 0,003.$$

Dans plusieurs applications, il est utile d'avoir une estimation de la queue de distribution $1 - \Phi(x)$. Nous avons le résultat suivant [6] :

Lemme 1.1. *Pour $x \rightarrow \infty$, alors*

$$1 - \Phi(x) \sim \frac{1}{x\sqrt{2\pi}} e^{-x^2/2};$$

plus précisément, pour tout $x > 0$

$$\left[\frac{1}{x} - \frac{1}{x^3}\right] \frac{1}{\sqrt{2\pi}} e^{-x^2/2} < 1 - \Phi(x) < \frac{1}{x} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

Le signe \sim est utilisé pour indiquer que le rapport des deux expressions tend vers 1.

Enfin pour terminer, indiquons une propriété importante de la loi normale, la stabilité par addition.

Proposition 1.2. *Soient X et Y deux v.a.r. indépendantes et de loi respectives, $\mathcal{N}(m_X, \sigma_X^2)$ et $\mathcal{N}(m_Y, \sigma_Y^2)$, alors la v.a.r. $X + Y$ est de loi $\mathcal{N}(m_X + m_Y, \sigma_X^2 + \sigma_Y^2)$.*

Pour prouver ce résultat, il suffit de multiplier les fonctions caractéristiques respectives. Il convient de retenir ce résultat très souvent utilisé dans les applications pratiques. Nous en verrons une extension ci-dessous.

2. Vecteurs aléatoires gaussiens

Définition 2.1. *Un vecteur aléatoire $X = (X_1, \dots, X_d)$, à valeurs dans $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, est dit gaussien si pour tout $\alpha = (\alpha_1, \dots, \alpha_d) \in \mathbb{R}^d$, la variable aléatoire réelle*

$$\langle \alpha, X \rangle = \sum_{i=1}^d \alpha_i X_i,$$

est gaussienne.

Un vecteur aléatoire est gaussien si et seulement si toutes combinaisons linéaires de ses composantes est une v.a.r. gaussienne.

La v.a.r. $\langle \alpha, X \rangle$, étant gaussienne, est entièrement caractérisée par sa moyenne

$$E\left(\sum_{1 \leq i \leq d} \alpha_i X_i\right) = \sum_{1 \leq i \leq d} \alpha_i E(X_i),$$

et sa variance

$$\text{Var}\left(\sum_{1 \leq i \leq d} \alpha_i X_i\right) = \sum_{1 \leq i, j \leq d} \alpha_i \alpha_j E((X_i - EX_i)(X_j - EX_j)).$$

Par conséquent, le vecteur gaussien $X = (X_1, \dots, X_d)$ est entièrement caractérisé par le vecteur moyenne

$$m = EX = (EX_1, \dots, EX_d)$$

et sa matrice de covariance

$$\Gamma = \left(E((X_i - EX_i)(X_j - EX_j)) \right)_{1 \leq i, j \leq d}.$$

La fonction caractéristique d'un vecteur gaussien est par définition

$$\varphi^X(t) = E(\exp(i\langle t, X \rangle)) = E \exp iY,$$

où $Y = \langle t, X \rangle$ est une v.a.r. gaussienne par hypothèse. Or la fonction caractéristique d'une v.a.r. gaussienne $\mathcal{N}(\mu, \sigma^2)$ est

$$\varphi^Y(t) = \exp(i\mu t - \frac{\sigma^2 t^2}{2}),$$

d'où

$$\varphi^X(t) = \exp(i\langle t, m \rangle - \frac{1}{2} t^T \Gamma t),$$

où t^T est le vecteur ligne transposé du vecteur colonne t .

Notons que si X est un vecteur gaussien, chacune de ses composantes X_i est une v.a.r. $\mathcal{N}(EX_i, \Gamma_{ii})$; il suffit de prendre $\alpha_j = \delta_{ij}$ dans la définition. Cependant, la réciproque est fautive, sauf si les X_i sont indépendantes. Par exemple, si Z est gaussienne réelle de loi $\mathcal{N}(0, 1)$ et si ϵ est de loi de Bernoulli symétrique indépendante de Z ($P\{\epsilon = 1\} = P\{\epsilon = -1\} = 1/2$), alors $(Z, \epsilon Z)$ n'est pas un vecteur gaussien, mais chacune des composantes suit une loi gaussienne.

D'autre part, si les composantes ne sont pas corrélées, donc si Γ est une matrice diagonale, alors elles sont indépendantes.

Proposition 2.2. *Soit X un vecteur gaussien de moyenne m et de matrice de covariance Γ . Sa fonction caractéristique est la suivante :*

$$\varphi^X(t) = E(\exp(i\langle t, X \rangle)) = \exp(i\langle t, m \rangle - \frac{1}{2}t^T \Gamma t).$$

De plus, les coordonnées d'un vecteur gaussien sont indépendantes si et seulement si elles ne sont pas corrélées.

Attention à ne pas appliquer l'équivalence entre indépendance et non corrélation à des variables gaussiennes quelconques. Reprenons l'exemple précédent, où Z et ϵZ sont des variables gaussiennes dépendantes mais non corrélées, puisque

$$E(Z\epsilon Z) = E(\epsilon Z^2) = E(\epsilon)E(Z^2) = 0 = E(\epsilon)E(Z).$$

2.1. Transformation linéaire d'un vecteur gaussien. Considérons un vecteur gaussien $G = (G_1, \dots, G_d)$ dont les composantes suivent une loi $\mathcal{N}(0, 1)$ et sont mutuellement indépendantes. Le vecteur G est alors centré et sa matrice de covariance est la matrice identité. La densité, par rapport à la mesure de Lebesgue sur \mathbb{R}^d , du vecteur G est

$$\phi(x) = \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{\|x\|^2}{2}\right).$$

Cet exemple, bien qu'élémentaire, est fondamental ; en effet rappelons qu'étant donné une matrice Γ symétrique semi-définie positive, il existe une unique matrice carrée A telle que $\Gamma = AA^T$. Ceci suggère qu'il doit être possible, de transformer tout vecteur gaussien en un vecteur du type G .

Proposition 2.3. *Soit $X = (X_1, \dots, X_d)$ un vecteur gaussien centré de matrice de covariance $\Gamma = AA^T$. Alors X a même loi que le vecteur AG où G est de loi $\mathcal{N}(0, Id)$. On notera $\mathcal{N}(0, \Gamma)$ la loi de X .*

DÉMONSTRATION. Les composantes du vecteur aléatoire AG s'exprimant comme une combinaison linéaire des composantes du vecteur G , nous en déduisons que AG est un vecteur gaussien. Notons $A = (a_{ij})_{1 \leq i, j \leq d}$. Pour tout $1 \leq i, j \leq d$

$$\begin{aligned} E((AG)_i(AG)_j) &= E\left(\left(\sum_{1 \leq k \leq d} a_{ik}G_k\right)\left(\sum_{1 \leq l \leq d} a_{jl}G_l\right)\right) \\ &= \sum_{1 \leq k \leq d} a_{ik}a_{jk} = \Gamma_{ij} \\ &= E(X_i X_j). \end{aligned}$$

La loi d'un vecteur gaussien étant déterminée par sa moyenne et sa matrice de covariance, on a bien prouvé que la loi de AG est la même que X . \square

On peut étendre le résultat précédent aux transformations linéaires.

Proposition 2.4. Soit $X = (X_1, \dots, X_d)$ un vecteur gaussien de loi $\mathcal{N}(m, \Gamma)$. Soient P est une matrice à k lignes et d colonnes et $a \in \mathbb{R}^k$. Considérons le vecteur aléatoire dans \mathbb{R}^k

$$Y = PX + a.$$

Alors Y est un vecteur gaussien de moyenne $Pm + a$ et de matrice de covariance $P\Gamma P^T$.

DÉMONSTRATION. On peut obtenir ce résultat à partir de la fonction caractéristique de Y (exercice). On peut aussi partir de la définition d'un vecteur gaussien, puisque

$$\begin{aligned} \langle \alpha, Y \rangle &= \langle \alpha, a \rangle + \langle P^T \alpha, X \rangle \\ &\stackrel{\mathcal{L}}{=} \langle \alpha, a \rangle + \mathcal{N}(\langle P^T \alpha, m \rangle, (P^T \alpha)^T \Gamma (P^T \alpha)) \\ &\stackrel{\mathcal{L}}{=} \mathcal{N}(\langle \alpha, Pm + a \rangle, \alpha^T (P\Gamma P^T) \alpha), \end{aligned}$$

où le symbole $\stackrel{\mathcal{L}}{=}$ signifie l'égalité en loi. On peut également utiliser la proposition précédente, car

$$\begin{aligned} X &\stackrel{\mathcal{L}}{=} m + AG \\ Y &\stackrel{\mathcal{L}}{=} a + m + (PA)G \\ &\stackrel{\mathcal{L}}{=} a + m + \mathcal{N}(0, P\Gamma P^T). \end{aligned}$$

□

2.2. Loi d'un vecteur gaussien non-dégénéré. Soit X un vecteur gaussien centré, de matrice de covariance $\Gamma = AA^T$, telle que A soit inversible. En vertu de la proposition précédente, le vecteur $A^{-1}X$ a même loi que G , par conséquent pour tout borélien B de \mathbb{R}^d , $P\{X \in B\} = P\{G \in A^{-1}(B)\}$. En appliquant la formule du changement de variable dans une intégrale, nous obtenons

$$\begin{aligned} P\{X \in B\} &= \frac{1}{(2\pi)^{d/2} |\det A|} \int_B \exp\left(-\frac{\|A^{-1}y\|^2}{2}\right) dy \\ &= \frac{1}{(2\pi)^{d/2} \sqrt{\det \Gamma}} \int_B \exp\left(-\frac{1}{2} y^T \Gamma^{-1} y\right) dy. \end{aligned}$$

Si la matrice Γ est dégénérée, on peut toujours l'écrire sous la forme $\Gamma = P\Delta P^T$, où P est une matrice orthogonale (c.-à-d. $P^{-1} = P^T$) et Δ une matrice diagonale positive, avec éventuellement des zéros sur la diagonale. Par exemple, un cas extrême est

$$\Delta = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

correspondant au vecteur gaussien $X = (X_1, X_2, X_3)$ où X_1 suit la loi $\mathcal{N}(0, 1)$, et X_2 et X_3 ont pour loi $\mathcal{N}(0, 0)$, c.-à-d. que $X_2 = X_3 = 0$ p.s.

La matrice Δ , étant positive, admet une racine carrée $\sqrt{\Delta}$, d'où

$$\Gamma = P\sqrt{\Delta}\sqrt{\Delta}P^T = (P\sqrt{\Delta})(P\sqrt{\Delta})^T,$$

et $A = P\sqrt{\Delta}$. Par conséquent, le vecteur aléatoire $P^T X$ est un vecteur gaussien de matrice de covariance la matrice diagonale Δ . En termes de changement de coordonnées, la diagonalisation de la matrice de covariance a permis de déterminer une nouvelle base dans laquelle les composantes de X sont orthogonales. On se ramène presque toujours à cette base. Comme Δ peut avoir des zéros sur la diagonale, le nombre de termes non nuls est en fait le rang de Γ ou le rang du vecteur gaussien X ; X est alors à valeurs dans un sous-espace vectoriel de \mathbb{R}^d de dimension le rang de X .

2.3. Ellipsoïde de dispersion. Soient $X \in \mathbb{R}^d$ un vecteur gaussien (centré) de matrice de covariance Γ inversible, et λ une constante positive. Considérons la surface \mathcal{E}_λ définie par

$$\mathcal{E}_\lambda = \{x \in \mathbb{R}^d : x^T \Gamma^{-1} x = \lambda^2\}.$$

Cette surface est une ellipsoïde, appelée ellipsoïde de dispersion. D'après la forme de la densité $f^X(x)$ du vecteur gaussien non-dégénéré X , l'ellipsoïde \mathcal{E}_λ correspond donc à la courbe de niveau

$$f^X(x) = \frac{1}{(2\pi)^{d/2} \sqrt{\det \Gamma}} e^{-\frac{\lambda^2}{2}}.$$

En utilisant la base des vecteurs propres de la matrice Γ , l'équation de l'ellipsoïde devient

$$\mathcal{E}_\lambda = \left\{x \in \mathbb{R}^d : \sum_{i=1}^d \left(\frac{x_i}{\sqrt{\lambda_i}}\right)^2 = \lambda^2\right\},$$

où les λ_i sont les valeurs propres de Γ . En particulier, pour $\lambda = 1$, l'ellipsoïde \mathcal{E}_1 , appelée ellipsoïde unité, a pour axes principaux les axes de coordonnées de la nouvelle base et pour demi-axes $\sqrt{\lambda_i}$.

Calculons maintenant la probabilité que le point correspondant au vecteur gaussien X tombe dans le volume D_λ délimité par l'ellipsoïde de dispersion \mathcal{E}_λ . Cette probabilité est le volume du cylindre délimité par la surface $z = f^X(x)$ et dont la base est le domaine D_λ . Nous avons

$$P\{X \in D_\lambda\} = \int_{D_\lambda} \frac{1}{(2\pi)^{d/2} \prod_i \lambda_i^{1/2}} \exp\left(-\frac{1}{2} \sum_{i=1}^d \left(\frac{x_i}{\sqrt{\lambda_i}}\right)^2\right) dx_1 \cdots dx_n.$$

Effectuons dans un premier temps le changement de coordonnées $v_i = x_i/\sqrt{\lambda_i}$. Le domaine D_λ devient la boule B_λ de \mathbb{R}^d de rayon λ , d'où

$$P\{X \in D_\lambda\} = \frac{1}{(2\pi)^{d/2}} \int_{B_\lambda} \exp\left(-\frac{\|x\|^2}{2}\right) dx.$$

Passons maintenant aux coordonnées sphériques (cf. 1-(15)), nous obtenons

$$P\{X \in D_\lambda\} = \frac{\sigma_d(1)}{(2\pi)^{d/2}} \int_0^\lambda r^{d-1} e^{-\frac{r^2}{2}} dr,$$

où $\sigma_n(1) = \frac{2\pi^{n/2}}{\Gamma(n/2)}$ est la surface de la sphère unité de dimension n . L'intégrale restante peut se calculer par récurrence. Par exemple, si $d = 2$, on obtient

$$P\{X \in D_\lambda\} = 1 - e^{-\frac{\lambda^2}{2}}.$$

Soit (X, Y) un couple de variables gaussiennes, dont le premier axe principal de l'ellipsoïde de dispersion unité fait un angle α avec l'axe Ox (voir figure 1). On supposera, pour simplifier, que l'ellipsoïde est centrée. Si λ_1 et λ_2 désignent respectivement la longueur des demi-axes, alors le vecteur de coordonnées (X, Y) peut s'écrire

$$\begin{cases} X &= \zeta \cos \alpha - \eta \sin \alpha \\ Y &= \zeta \sin \alpha + \eta \cos \alpha, \end{cases}$$

où ζ est une v.a.r. de loi $\mathcal{N}(0, \lambda_1^2)$, η de loi $\mathcal{N}(0, \lambda_2^2)$ et sont indépendantes l'une de l'autre (ζ représente l'erreur dans le sens du plus grand axe et η l'erreur dans le sens transversale).

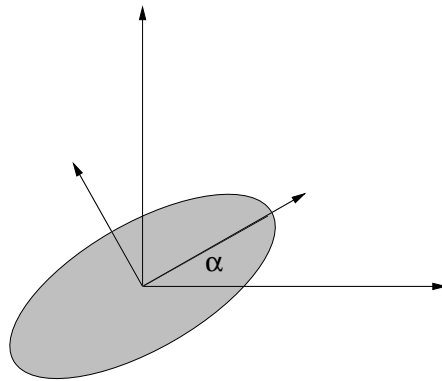


FIG. 1. Couple de v.a. gaussiennes

Par identification, on en déduit les relations suivantes :

$$\begin{aligned} \sigma_X^2 &= \lambda_1^2 \cos^2 \alpha + \lambda_2^2 \sin^2 \alpha \\ \sigma_Y^2 &= \lambda_1^2 \sin^2 \alpha + \lambda_2^2 \cos^2 \alpha \\ \tan(2\alpha) &= 2 \frac{\text{Cov}(X, Y)}{\sigma_X^2 - \sigma_Y^2} \end{aligned}$$

Exemple 53. Soit A_1 un avion suivant une trajectoire rectiligne dans le plan. Supposons que les erreurs de navigations selon les axes Ox et Oy suivent une loi gaussienne centrée de variance respective σ_x^2 et σ_y^2 . Pour simplifier, nous supposons de plus que les erreurs transversales et longitudinales sont

indépendantes. Par conséquent, les ellipsoïdes d'inertie ont pour axes principaux, l'axe de la trajectoire et l'axe transversal. Le vecteur modélisant la position relative de l'avion A_1 par rapport à un autre avion A_2 est également un vecteur gaussien, mais la présence de corrélations possibles entre les erreurs pour chaque avion (dues par exemple au vent) conduit à une matrice de covariance qui n'est plus diagonale.

On s'intéresse ensuite à déterminer la probabilité de conflit entre ces deux avions supposés suffisamment proches. Pour cela on définit un volume de conflit autour de chaque avion, ici un cercle, et on dira qu'il y a conflit si un des avions pénètre dans le volume de conflit de l'autre avion. En utilisant les positions relatives des deux avions, la zone de conflit devient un cercle de rayon, la somme des rayons de chacune des zones de conflit définie pour chaque avion. La probabilité qu'il y ait conflit à un instant fixé, sera donc donné par le volume délimité par le cylindre, dont la base est l'intersection de la zone de conflit et de l'ellipsoïde \mathcal{E}_λ (celle-ci déterminant la position de l'avion A_1 , relativement à l'avion A_2 , avec une probabilité de $1 - \exp(-\lambda^2/2)$). Ceci est le point de départ de l'article [12].

Exemple 54. [14] Un parachutiste saute d'un aérostat et essaie d'atterrir en un certain point. L'erreur du point d'atterrissage se décompose en

1. l'erreur liée à ce que la direction du vent et la vitesse du vent sont différentes aux différentes altitudes,
2. des erreurs dues aux autres causes.

Ces deux groupes d'erreurs sont indépendants. Les axes principaux de la dispersion due au vent sont dirigés dans la direction du vent et perpendiculairement. Les écarts types sont respectivement 20 et 10 m. Les erreurs dues aux autres causes ont pour ellipse de dispersion unitaire, un cercle de rayon égal à 15 m. Les erreurs systématiques sont exclues, donc toutes les variables sont centrées. Trouvons les paramètres de la loi normale que suit l'erreur totale du point d'atterrissage.

Pour cela, choisissons pour axe Ox la direction dominante du vent et pour axe Oy la perpendiculaire. En appliquant la règle de composition des lois normales à composantes indépendantes, on obtient

$$\begin{aligned} m_x &= m_y = 0, \\ \sigma_x^2 &= 20^2 + 15^2 = 625, \\ \sigma_y^2 &= 10^2 + 15^2 = 325. \end{aligned}$$

Exemple 55. [14] Pour faire son point, un avion mesure la distance le séparant de deux repères se trouvant à 50 km à l'ouest et à 100 km au sud-ouest. Les erreurs de mesure sont indépendantes gaussiennes centrées d'écart type égal à 100 m. Trouver les paramètres de la loi de répartition de l'erreur de position de l'avion.

Prenons pour axe Ox , l'axe dirigé vers l'est et l'axe Oy dirigé vers le nord. Désignons respectivement par (a_1, b_1) et (a_2, b_2) les coordonnées de

deux repères. La position de l'avion est représentée par le point (x, y) et les distances vraies aux repères sont respectivement

$$\sqrt{(x - a_1)^2 + (y - b_1)^2}, \quad \sqrt{(x - a_2)^2 + (y - b_2)^2}.$$

En raison des erreurs, les distances mesurées peuvent s'exprimer par

$$\begin{aligned} R_1 &= \sqrt{(x - a_1)^2 + (y - b_1)^2} + \Delta R_1, \\ R_2 &= \sqrt{(x - a_2)^2 + (y - b_2)^2} + \Delta R_2, \end{aligned}$$

où ΔR_1 et ΔR_2 sont des v.a.r. indépendantes. La position mesurée de l'avion (X, Y) s'obtient à partir du système d'équations

$$\begin{cases} (X - a_1)^2 + (Y - b_1)^2 = R_1^2 \\ (X - a_2)^2 + (Y - b_2)^2 = R_2^2 \end{cases}$$

Cherchons une solution de la forme $X = x + \Delta X$ et $Y = y + \Delta Y$. On obtient après linéarisation

$$\begin{cases} 2(x - a_1)\Delta X + 2(y - b_1)\Delta Y \approx 2\sqrt{(x - a_1)^2 + (y - b_1)^2}\Delta R_1 \\ 2(x - a_2)\Delta X + 2(y - b_2)\Delta Y \approx 2\sqrt{(x - a_2)^2 + (y - b_2)^2}\Delta R_2. \end{cases}$$

Comme

$$\begin{aligned} x - a_1 &= 50 \text{ km}, & y - b_1 &= 0, \\ x - a_2 &= y - b_2 = 50\sqrt{2} \text{ km} \end{aligned}$$

on obtient que

$$\Delta X \approx \Delta R_1, \quad \Delta Y \approx \sqrt{2}\Delta R_2 - \Delta R_1.$$

On obtient ainsi

$$\begin{aligned} m_{\Delta X} &= m_{\Delta Y} = 0, \\ \sigma_{\Delta X}^2 &= \sigma_{\Delta R_1}^2 = 100^2, \\ \sigma_{\Delta Y}^2 &= 2\sigma_{\Delta R_2}^2 + \sigma_{\Delta R_1}^2 = 30000, \\ \text{Cov}(\Delta X, \Delta Y) &= -\sigma_{\Delta R_1}^2 = -10000. \end{aligned}$$

Supposons maintenant que cet avion jette sur terre une caisse. La dispersion du point d'atterrissage de la caisse, due à des causes techniques, est circulaire d'écart type 70 m et l'erreur systématique étant de 30 m vers le nord (direction du vol de l'avion). Trouver les paramètres de la loi de répartition de l'erreur $(\Delta X, \Delta Y)$ du point d'atterrissage due aux causes techniques et à l'imprécision de détermination de la position de l'avion.

Désignons par $(\Delta X_1, \Delta Y_1)$ l'erreur due à l'imprécision de la position de l'avion et par $(\Delta X_2, \Delta Y_2)$ celle due aux causes techniques. Les paramètres

des composantes sont :

$$\begin{aligned} m_{\Delta X_1} &= m_{\Delta Y_1} = m_{\Delta X_2} = 0, & m_{\Delta Y_2} &= 30, \\ \sigma_{\Delta X_1}^2 &= 10000, & \sigma_{\Delta Y_1}^2 &= 30000, & \text{Cov}(\Delta X_1, \Delta Y_1) &= -10000, \\ \sigma_{\Delta X_2}^2 &= \sigma_{\Delta Y_2}^2 = 4900, & \text{Cov}(\Delta X_1, \Delta Y_1) &= 0. \end{aligned}$$

Les erreurs étant indépendantes, on obtient :

$$\begin{aligned} m_{\Delta X} &= 0, & m_{\Delta Y} &= 30, \\ \sigma_{\Delta X}^2 &= 14900, & \sigma_{\Delta Y}^2 &= 34900, \\ \text{Cov}(\Delta X, \Delta Y) &= -10000. \end{aligned}$$

En particulier, l'angle α formé par l'axe Ox et le premier axe principal de dispersion est

$$\tan(2\alpha) = \frac{2 \text{Cov}(\Delta X, \Delta Y)}{\sigma_{\Delta X}^2 - \sigma_{\Delta Y}^2} = 1,$$

soit $\alpha = 22^\circ 30'$. Les longueurs des demi-axes s'expriment par

$$\begin{aligned} \lambda_1^2 &= \sigma_{\Delta X}^2 \cos^2 \alpha + \sigma_{\Delta Y}^2 \sin^2 \alpha + \text{Cov}(\Delta X, \Delta Y) \sin(2\alpha) \approx 104^2 \\ \lambda_2^2 &= \sigma_{\Delta X}^2 \sin^2 \alpha + \sigma_{\Delta Y}^2 \cos^2 \alpha - \text{Cov}(\Delta X, \Delta Y) \sin(2\alpha) \approx 198^2. \end{aligned}$$

3. Espérances conditionnelles et variables gaussiennes

Soient X une variable aléatoire réelle et $Y = (Y_1, \dots, Y_d)$ un vecteur aléatoire. Nous avons vu, au §3-(4), que $E(X|Y)$ était la meilleure approximation quadratique de X parmi les variables de la forme $f(Y)$, où f est une fonction borélienne de \mathbb{R}^d dans \mathbb{R} . Dans le cas des variables gaussiennes, la meilleure approximation quadratique est obtenue lorsque f est une fonction affine.

Théorème 3.1. *Supposons que le vecteur aléatoire (X, Y_1, \dots, Y_d) de \mathbb{R}^{d+1} soit un vecteur gaussien. Alors*

(i) *il existe des constantes a, b_1, \dots, b_d , telles que*

$$E(X|Y) = a + \sum_{i=1}^d b_i Y_i,$$

(ii) *il existe une variable aléatoire gaussienne réelle Z centrée indépendante de Y telle que*

$$X = a + \sum_{i=1}^d b_i Y_i + Z,$$

(iii) *la loi conditionnelle de X sachant $Y = (y_1, \dots, y_d)$ est gaussienne.*

DÉMONSTRATION. [3] Le sous-espace $H \in L^2(\Omega, \mathcal{A}, P)$ des variables de la forme

$$\alpha + \sum_{i=1}^d \beta_i Y_i$$

est fermé (car de dimension finie). Par conséquent, la projection orthogonale de X sur H définit un unique vecteur X_0 tel que

$$X_0 \in H, \quad \text{et} \quad X - X_0 \perp H.$$

Par conséquent, il existe a, b_1, \dots, b_d tels que

$$X_0 = a + \sum_{i=1}^d b_i Y_i.$$

Soit $Z = X - X_0$, alors Z étant perpendiculaire à H qui contient la v.a.r. constante égale à 1, on en déduit que $E(Z1) = E(Z) = 0$, donc que Z est centré. D'autre part, en tant que combinaison de v.a.r. gaussiennes, Z est une v.a.r. gaussienne et $E(ZY_i) = 0 = E(Z)E(Y_i)$. Or, nous avons vu que dans le cas des variables gaussiennes, la non corrélation implique l'indépendance, d'où Z est indépendant de tous les Y_i , donc du vecteur Y .

Il en résulte que

$$E(X|Y) = E(Z|Y) + E(X_0|Y) = E(Z) + X_0 = a + \sum_{i=1}^d b_i Y_i,$$

ce qui démontre (i) et (ii). Pour prouver (iii), utilisons la fonction caractéristique conditionnelle. Il vient

$$\begin{aligned} E(e^{itX}|Y) &= e^{iE(tX|Y)} E\left(e^{i(tX - E(tX|Y))}|Y\right) \\ &= e^{iE(tX|Y)} E\left(e^{itZ}|Y\right) \\ &= \exp\left(iE(tX|Y) - \frac{t^2}{2}E(Z^2)\right), \end{aligned}$$

où nous avons utilisé l'indépendance de Z et Y et donc que $E(\phi(Z)|Y) = E(\phi(Z))$, pour toute fonction mesurable ϕ . Or comme

$$E(e^{itX}|Y) = \int e^{itx} K^Y(dx),$$

on en déduit que la loi conditionnelle de X sachant $Y = y$ est une loi gaussienne de moyenne $E(X|Y = y) = a + \sum_i b_i y_i$ et de variance $E(X - E(X|Y = y))^2$. \square

Ces résultats s'étendent au cas où $(X_1, \dots, X_k, Y_1, \dots, Y_d)$ est un vecteur gaussien à valeurs dans \mathbb{R}^{k+d} . Alors la loi conditionnelle de X sachant $Y = y$ est une loi gaussienne de moyenne le vecteur (de \mathbb{R}^k) $E(X|Y = y)$ et de matrice de covariance $E((X_i - E(X_i|Y))(X_j - E(X_j|Y)))$.

3.1. Application pratique. Comment utiliser les résultats précédents dans la pratique? Supposons donc que $(X_1, \dots, X_k, Y_1, \dots, Y_d)$ soit un vecteur gaussien dans \mathbb{R}^{k+d} . Alors, pour tout $1 \leq i \leq k$, nous avons

$$E(X_i|Y) = a_i + \sum_{j=1}^d b_{ij} Y_j.$$

Pour déterminer a_i et les b_{ij} , il suffit de calculer les espérances respectives de X_i et des $Y_k X_i$. En effet,

$$E(X_i) = E(E(X_i|Y)) = a_i + \sum_{j=1}^d b_{ij} E(Y_j)$$

$$E(Y_k X_i) = E(Y_k E(X_i|Y)) = a_i E(Y_k) + \sum_{j=1}^d b_{ij} E(Y_k Y_j).$$

Nous obtenons ainsi un système linéaire de $d + 1$ équations à $d + 1$ inconnues. En particulier si (X, Y) est un vecteur gaussien centré de matrice de covariance

$$C = \begin{pmatrix} C_X & D \\ D^T & C_Y \end{pmatrix},$$

alors $a_i = 0$ pour tout $1 \leq i \leq k$ et, en introduisant la matrice $(k \times d)$ $B = (b_{ij})$, nous avons $B = D^T C_Y^{-1}$.

Exemple 56. Soit (X, Y_1, Y_2) un vecteur gaussien centré de matrice de covariance

$$\begin{pmatrix} 5 & 0 & 3 \\ 0 & 1 & -1 \\ 3 & -1 & 4 \end{pmatrix}.$$

Calculons $E(X|Y_1, Y_2)$. Pour cela, écrivons que $E(X|Y_1, Y_2) = b_1 Y_1 + b_2 Y_2$, puis

$$E(X Y_1) = b_1 E(Y_1^2) + b_2 E(Y_2 Y_1)$$

$$E(X Y_2) = b_1 E(Y_1 Y_2) + b_2 E(Y_2^2)$$

ce qui conduit au système d'équations suivant

$$\begin{cases} 0 & = b_1 - b_2 \\ 3 & = -b_1 + 4b_2 \end{cases}.$$

La solution est $b_1 = b_2 = 1$, par conséquent $E(X|Y_1, Y_2) = Y_1 + Y_2$.

Convergence de suites de variables aléatoires

Nous abordons dans ce chapitre, les questions relatives au comportement asymptotique de suites de variables aléatoires. Nous y verrons en particulier, l'apport de l'axiomatisation de la théorie des probabilités au moyen de la théorie de la mesure.

Il existe de nombreuses notions de convergence de variables aléatoires qui sont essentielles dans les applications. Elles servent surtout à montrer que les phénomènes aléatoires présentent certaines régularités permettant d'identifier certaines de leurs propriétés. Il sera alors possible d'avoir une quasi-certitude sur des phénomènes aléatoires. Nous avons déjà rencontré un résultat de convergence, à savoir la loi faible des grands nombres pour une suite de v.a.r. indépendantes et de même loi 3-(6). La première démonstration rigoureuse de la loi faible des grands nombres date de 1713, lors de la parution à titre posthume du mémoire de J. Bernoulli, *L'art de la conjecture*. Il y a en effet prouvé que si X_1, \dots, X_n, \dots est une suite de v.a.r. indépendantes pouvant chacune prendre la valeur 1 avec probabilité p et la valeur 0 avec probabilité $1 - p$, alors si

$$S_n = X_1 + \dots + X_n$$

désigne le nombre de fois que X_i a pris la valeur 1, pour $1 \leq i \leq n$, alors, pour tout $\epsilon > 0$ fixé, la probabilité $P(|S_n/n - p| > \epsilon)$ tend vers 0 lorsque n tend vers l'infini. La loi des grands nombres de Bernoulli ne précise pas la vitesse de convergence vers 0. Ce problème ne fût résolu par de Moivre qu'en 1718, prouvant que, pour tout $-\infty < a < b < \infty$ fixés (avec $p \neq 0$ ou 1) :

$$\lim_{n \rightarrow \infty} P\left(a < \frac{S_n - np}{\sqrt{np(1-p)}} < b\right) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-t^2/2} dt.$$

On en déduit, en particulier

$$\begin{aligned} P\left(\left|\frac{S_n}{n} - p\right| > \epsilon\right) &= 1 - P\left(-\frac{n^{1/2}\epsilon}{\sqrt{np(1-p)}} < \frac{S_n - np}{\sqrt{np(1-p)}} < \frac{n^{1/2}\epsilon}{\sqrt{np(1-p)}}\right) \\ &\approx 2\left(1 - \Phi\left(\frac{n^{1/2}\epsilon}{\sqrt{np(1-p)}}\right)\right) \\ &\approx \frac{\sqrt{2p(1-p)}}{\sqrt{n\pi}\epsilon} \exp\left(-\frac{n\epsilon}{2p(1-p)}\right). \end{aligned}$$

Il convient de noter l'insuffisance de la loi faible des grands nombres, puisqu'elle nous dit que pour n grand, le rapport S_n/n est proche de p avec une

grande probabilité. Il serait beaucoup plus souhaitable d'obtenir un résultat de la forme

$$(0.1) \quad \frac{S_n}{n} \rightarrow p.$$

Mais un tel résultat est faux en général. Considérons par exemple, le jet d'une pièce et notons 1 si face sort et 0 sinon. S_n/n est donc la fréquence relative d'apparition de face dans n jets. Soit Ω l'ensemble des suites $(u_n)_{n \geq 1}$, avec $u_i = 1$, ou 0 (signalons que Ω a même cardinal que l'intervalle $[0, 1]$). La relation 0.1 signifie que pour toute suite (u_n)

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n u_i = p.$$

Considérons la suite suivante [3]

$$\begin{aligned} u_1 &= 1, u_2 = 0, u_3 = 0 \\ u_n &= 1, \text{ si } 3^{2p-1} < n \leq 3^{2p} \\ u_n &= 0, \text{ si } 3^{2p} < n \leq 3^{2p+1}, \end{aligned}$$

alors l'expression $n^{-1} \sum_{i=1}^n u_i$ n'a pas de limite.

Le mieux que l'on puisse espérer est que la probabilité de l'ensemble des suites ne vérifiant pas la relation 0.1 soit de probabilité nulle. Cette propriété se dénomme loi forte des grands nombres et fait intervenir la notion de convergence presque sûre. Nous verrons d'autres types de convergence, en particulier la convergence en loi, dont nous avons déjà vu des exemples lorsque nous avons calculé la limite, quand n tend vers l'infini, d'une suite de fonctions de répartition (F_n) . Le résultat le plus important concernant ce type de limite est le théorème central limite.

Dans tout ce chapitre, les suites de variables aléatoires réelles $(X_n)_{n \in \mathbb{N}}$ sont supposées construites sur le même espace de probabilité (Ω, \mathcal{A}, P) .

1. Convergence presque sûre

Définition 1.1. *On dit que la suite de v.a.r. (X_n) converge vers la variable aléatoire X presque sûrement, si il existe un ensemble P -négligeable N (c.-à-d. $P(N) = 0$) tel que*

$$\forall \omega \notin N, \quad X_n(\omega) \xrightarrow[n \rightarrow \infty]{} X,$$

ou en d'autre termes

$$P\{\omega \in \Omega : \lim_{n \rightarrow \infty} X_n(\omega) = X(\omega)\} = 1.$$

Dans ce cas, on note $\lim_{n \rightarrow \infty} X_n = X$ p.s. ou $X_n \rightarrow X$ p.s. lorsque $n \rightarrow \infty$.

Notons que l'événement $\{\lim_{n \rightarrow \infty} X_n = X\}$ est bien mesurable puisqu'égal à

$$\bigcap_{p \geq 1} \bigcup_{m \in \mathbb{N}} \bigcap_{n \geq m} \{|X_n - X| < \frac{1}{p}\}.$$

Or si $(A_p)_{p \in \mathbb{N}}$ est une suite d'événements, $P(\cap_{p \in \mathbb{N}} A_p) = 1$ si et seulement si $P(A_p) = 1$ pour tout p . Il s'ensuit que X_n converge vers X p.s. si et seulement si

$$\forall \epsilon > 0, \quad P\left(\bigcup_{m \in \mathbb{N}} \bigcap_{n \geq m} \{|X_n - X| < \epsilon\}\right) = 1.$$

Cette condition peut aussi s'écrire, par passage au complémentaire,

$$\forall \epsilon > 0, \quad P\left(\bigcap_{m \in \mathbb{N}} \bigcup_{n \geq m} \{|X_n - X| \geq \epsilon\}\right) = 0,$$

soit, en utilisant la notation du § 1-(5.3), la condition équivalente

$$(1.1) \quad \forall \epsilon > 0, \quad P\{|X_n - X| \geq \epsilon \text{ i.s.}\} = 0.$$

Notons que les ensembles

$$A_m := \bigcup_{n \geq m} \{|X_n - X| < \epsilon\}, \quad m \in \mathbb{N}$$

sont décroissants, donc par convergence monotone, nous avons

$$P\left(\bigcap_{m \in \mathbb{N}} \bigcup_{n \geq m} \{|X_n - X| \geq \epsilon\}\right) = \lim_{m \rightarrow \infty} P\{\sup_{n \geq m} |X_n - X| \geq \epsilon\},$$

d'où X_n converge vers X p.s. si et seulement si

$$(1.2) \quad \forall \epsilon > 0, \quad \lim_{m \rightarrow \infty} P\{\sup_{n \geq m} |X_n - X| \geq \epsilon\} = 0.$$

Un des moyens les plus courant pour prouver la convergence p.s. est le lemme de Borel-Cantelli.

Lemme 1.2. *Soit $(X_n)_{n \in \mathbb{N}}$ une suite de variables aléatoires réelles définies sur (Ω, \mathcal{A}, P) .*

- (i) *Si pour tout $\epsilon > 0$, $\sum_{n \in \mathbb{N}} P\{|X_n - X| > \epsilon\} < \infty$, alors X_n converge vers X p.s.,*
- (ii) *si les (X_n) sont mutuellement indépendantes, alors X_n converge vers 0 p.s. si et seulement si la série $\sum_{n \in \mathbb{N}} P\{|X_n| > \epsilon\}$ converge pour tout $\epsilon > 0$.*

DÉMONSTRATION. Pour prouver (i), nous allons utiliser le lemme de Borel-Cantelli exprimé pour une famille d'ensemble (cf. § 1-(5.3)). En effet, pour $\epsilon > 0$ considérons les événements

$$A_n = \{|X_n - X| \geq \epsilon\}, \quad n \in \mathbb{N}.$$

Appliquons le lemme de Borel-Cantelli à ces ensembles, il vient $P(A_n \text{ i.s.}) = 0$ ce qui fournit le résultat d'après 1.1.

Pour prouver (ii), on utilisera la réciproque du lemme de Borel-Cantelli du § 1-(5.3). Pour cela, il faut que la famille des (A_n) soit mutuellement indépendante. Or si X est une constante, A_n appartient à la tribu engendrée par X_n , et donc l'hypothèse d'indépendance mutuelle des (X_n) implique celle des A_n . \square

Exemple 57. Soit une suite de v.a.r. (X_n) telle que pour un $p > 0$,

$$\sum_{n=1}^{\infty} E(|X_n|^p) < \infty.$$

Alors d'après l'inégalité de Tchebychev

$$P\{|X_n| > \epsilon\} \leq \epsilon^{-p} E(|X_n|^p),$$

par conséquent, la série $\sum_n P\{|X_n| > \epsilon\}$ converge, d'où X_n converge vers X p.s.

Un critère de convergence presque sûre est le critère de Cauchy. A savoir que X_n converge vers X p.s. si et seulement si

$$\forall \epsilon > 0, \quad P\left(\bigcup_{m \in \mathbb{N}} \bigcap_{n \geq m} \{|X_n - X_m| \leq \epsilon\}\right) = 1.$$

Pour cela, il suffit d'appliquer le critère de Cauchy pour les suites de nombres réels en remarquant que (X_n) (resp. $X_n - X_m$, $n, m \geq 1$) ne converge p.s. que si les suites $(X_n(\omega))$ (resp. $X_n(\omega) - X_m(\omega)$, $n, m \geq 1$) convergent dans \mathbb{R} pour tout $\omega \notin N$.

Exemple 58. Soit $(X_n)_{n \in \mathbb{N}}$ une suite de v.a. indépendantes et de même loi de Bernoulli $\mathcal{B}(p)$, c.-à-d. que $P\{X_i = 1\} = 1 - P\{X_i = 0\} = p$. Soit $U_n = \sum_{1 \leq i \leq n} 2^{-i} X_i$. Pour montrer la convergence p.s. de U_n , on peut appliquer le critère de Cauchy, en notant que $n < m$ implique

$$|U_n - U_m| \leq \sum_{n+1 \leq i \leq m} 2^{-i} \leq 2^{-n}.$$

Ainsi,

$$\begin{aligned} \bigcup_{n \in \mathbb{N}} \bigcap_{m \geq n} \{\omega : |U_n(\omega) - U_m(\omega)| \leq \epsilon\} &\supset \bigcup_{n \in \mathbb{N}} \bigcap_{m \geq n} \{\omega : 2^{-n} \leq \epsilon\} \\ &= \bigcup_{n \in \mathbb{N}} \{\omega : 2^{-n} \leq \epsilon\} \\ &= \Omega. \end{aligned}$$

2. Convergence en probabilité

Définition 2.1. On dit que la suite (X_n) converge en probabilité vers la variable aléatoire X , si pour tout $\epsilon > 0$

$$\lim_{n \rightarrow \infty} P\{|X_n - X| \geq \epsilon\} = 0.$$

On notera $X_n \xrightarrow{P} X$, ou $P - \lim_{n \rightarrow \infty} X_n = X$ cette propriété de convergence.

La convergence en probabilité est une propriété plus faible que la convergence presque sûre. La convergence presque sûre exige en effet, la convergence en probabilité des suprema $Y_n = \sup_{m \geq n} |X_m - X|$, d'après 1.2.

Exemple 59. Soit $(X_i)_{i \geq 1}$ une suite de v.a.r. non corrélées, telles que $E(X_i) = 0$ et $\text{Var}(X_i) = \sigma^2$. Alors leurs moyennes partielles (ou empirique) $n^{-1} \sum_{1 \leq i \leq n} X_i$ convergent en probabilité vers 0. En effet, d'après l'inégalité de Tchébycheff

$$P\left\{\left|\frac{1}{n} \sum_{1 \leq i \leq n} X_i\right| \geq \epsilon\right\} \leq \frac{1}{n^2 \epsilon^2} \text{Var}\left(\sum_{1 \leq i \leq n} X_i\right) = \frac{\sigma^2}{n \epsilon^2}.$$

Exemple 60. Soit $(X_n)_{n \geq 1}$ une suite de variables aléatoires indépendantes, de loi de Bernoulli avec $P\{X_n = 1\} = 1 - P\{X_n = 0\} = p_n$. Alors

$$X_n \xrightarrow{P} 0 \iff \lim_{n \rightarrow \infty} p_n = 0.$$

Mais le lemme de Borel-Cantelli montre que

$$\lim_{n \rightarrow \infty} X_n = 0 \text{ p.s.} \iff \sum_n P\{|X_n| > \epsilon\} < \infty \text{ pour tout } \epsilon > 0.$$

Ceci est équivalent à $\sum_n p_n < \infty$. En particulier, si $p_n = 1/n$ on a une convergence en probabilité vers 0, mais pas presque sûrement.

Ce dernier exemple prouve bien que la convergence en probabilité n'entraîne pas la convergence presque sûre. Cependant, nous avons le résultat important suivant, que nous admettrons :

Proposition 2.2. *Soit $(X_n)_{n \geq 1}$ une suite de variables aléatoires réelles. Alors X_n converge en probabilité vers X , si et seulement si de toute suite croissante d'entiers (n') on peut extraire une sous-suite (n'') telle que $X_{n''}$ converge vers X presque sûrement.*

La convergence en probabilité présente une certaine stabilité pour les opérations algébriques.

Proposition 2.3. *Soient $(X_n)_{n \geq 1}$ et $(Y_n)_{n \geq 1}$ deux suites de variables aléatoires réelles définies sur un espace (Ω, \mathcal{A}, P) . Supposons que X_n (resp. Y_n) converge en probabilité vers une variable aléatoire X (resp. Y) définie sur (Ω, \mathcal{A}, P) .*

- (i) *Si φ est une application continue de \mathbb{R} dans \mathbb{R} , alors $\varphi(X_n) \xrightarrow{P} \varphi(X)$.*
- (ii) *Pour tout $\alpha, \beta \in \mathbb{R}$,*

$$\alpha X_n + \beta Y_n \xrightarrow{P} \alpha X + \beta Y.$$

- (iii) *Soit Z une v.a.r. finie p.s., alors $ZX_n \xrightarrow{P} ZX$.*

DÉMONSTRATION. Vérifions par exemple (ii). Soit (n') une suite partielle. On peut en extraire une sous-suite (n'') telle que $X_{n''}$ converge vers X p.s. De (n'') , on peut extraire une sous-suite (n''') telle que $Y_{n'''}$ converge vers Y p.s. Alors $\alpha X_{n'''} + \beta Y_{n'''}$ converge vers $\alpha X + \beta Y$ p.s. Il suffit de conclure d'après la proposition 2.2. \square

En fait, la convergence en probabilité est métrisable, il suffit de poser

$$d(X, Y) = E\left(\frac{|X - Y|}{1 + |X - Y|}\right).$$

On montre que l'espace des classes d'équivalence de v.a.r. muni de cette mesure est complet [1].

Les propriétés précédentes et la proposition 2.2 découlent de cette propriété. Notons que la convergence presque sûre n'est pas métrisable, car de la proposition 2.2, il en découlerait la coïncidence des deux notions de convergence.

3. Convergence dans les espaces L^p

Rappelons qu'une variable aléatoire réelle X , définie sur (Ω, \mathcal{A}, P) est dans $L^p(\Omega, \mathcal{A}, P)$, pour $1 \leq p < \infty$, si et seulement si $E(|X|^p) < \infty$. L'espace $L^p(\Omega, \mathcal{A}, P)$, muni de la norme

$$\|X\|_p = (E(|X|^p))^{1/p},$$

est un espace métrique complet (espace de Banach). On peut donc définir une notion de convergence dans ces espaces.

Rappelons également que si $X \in L^q$, alors $X \in L^p$ pour tout $p \leq q$ (cf (iii) du théorème 2-(3.7)).

Définition 3.1. Soit $(X_n)_{n \in \mathbb{N}}$ une suite de variables aléatoires réelles dans $L^p(\Omega, \mathcal{A}, P)$. On dira que X_n converge vers une variable aléatoire X dans L^p , si $X \in L^p(\Omega, \mathcal{A}, P)$ et si

$$\lim_{n \rightarrow \infty} \|X_n - X\| = 0,$$

ou de façon équivalente si

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

Si $p = 2$, on parlera de convergence en moyenne quadratique.

Remarque 3.2. Si (X_n) est une suite de v.a.r. intégrables (c.-à-d. dans L^1), alors la convergence presque sûre vers X n'implique pas la convergence dans L^1 . Par exemple, soit $\Omega = \mathbb{R}$ muni de sa tribu borélienne. Soit (X_n) une suite de v.a.r. telles que $P\{X_n = n\} = n^{-p} = 1 - P\{X_n = 0\}$, avec $p > 1$. Si $\epsilon > 0$, alors pour tout $n \geq 1/\epsilon^p$, $P\{|X_n| \geq \epsilon\} = n^{-p}$. Comme $p > 1$, la série correspondante converge et d'après le lemme de Borel-Cantelli, on en déduit que X_n converge p.s. vers 0. En revanche, $E(|X_n|^p) = 1$ pour tout n .

De même, la convergence en probabilité n'entraîne pas la convergence dans L^p . Il suffit de considérer l'exemple précédent avec $p = 1$ et de supposer les variables aléatoires indépendantes. On a alors convergence en probabilité vers 0, mais ni convergence presque sûre, ni convergence dans L^1 .

Remarque 3.3. Inversement, si X_n converge dans L^p , cela n'entraîne pas nécessairement la convergence presque sûre. Par contre, la convergence dans

L^1 implique la convergence en probabilité. Il suffit d'appliquer l'inégalité de Markov.

Remarque 3.4. D'après le rappel fait au début de cette section, si X_n converge vers X dans L^q , alors X_n converge vers X dans tous les L^p , avec $1 \leq p \leq q$.

4. Convergence en loi

Ce quatrième type de convergence est le plus faible de tous, et le plus souvent utilisé dans les applications.

Nous avons vu que deux variables aléatoires X et Y définies sur le même espace de probabilité (Ω, \mathcal{A}, P) ont même lois si et seulement si

$$F^X = F^Y \quad (\text{égalité des fonctions de répartition})$$

ou

$$E(f(X)) = E(f(Y)) \quad \text{pour toute fonction continue bornée } f$$

ou

$$\varphi^X = \varphi^Y \quad (\text{égalité des fonctions caractéristiques})$$

Ceci conduit à la notion de convergence en loi suivante.

Définition 4.1. Soient (X_n) une suite de variables aléatoires réelles définies sur (Ω, \mathcal{A}, P) et soit X une variable aléatoire réelle définie sur un espace de probabilité (pouvant être différent du précédent). On dit que X_n converge en loi vers X , si l'une des trois conditions équivalentes suivantes est vérifiée :

- (i) $\lim_{n \rightarrow \infty} F^{X_n}(t) = F^X(t)$ en tout point de continuité de F^X ,
- (ii) $\lim_{n \rightarrow \infty} \int f(X_n) dP = \int f(X) dP$, pour toute fonction continue bornée f de \mathbb{R} dans \mathbb{R} ,
- (iii) $\lim_{n \rightarrow \infty} \varphi^{X_n}(t) = \varphi^X(t)$, pour tout $t \in \mathbb{R}$.

Signalons que cette définition est en fait une définition et un théorème, puisqu'il conviendrait de prouver l'équivalence des quatre conditions. Pour cela voir [1] ou [3]. Notons également que la terminologie employée peut présenter un risque. En effet, ce ne sont pas les variables aléatoires qui convergent mais leurs lois. La variable aléatoire X n'est pas définie de façon unique, mais seulement par sa loi.

Seule la convergence presque sûre autorise les opérations algébriques. Il en est de même pour certaines opérations algébriques, dans le cas de la convergence en probabilité d'après la proposition 2.3. Par exemple, si X_n converge en loi vers X , définie sur le même espace de probabilité, cela n'implique pas que $X_n - X$ converge vers 0 en loi. En effet, prenons X de loi uniforme sur l'intervalle $[-1, 1]$, cette loi étant symétrique par rapport à l'origine, la suite $X_n = -X$, pour tout n , converge en loi vers X , mais $X_n - X = -2X$ converge en loi vers $-2X$ de loi uniforme sur l'intervalle $[-2, 2]$.

Remarque 4.2. Nous avons vu en introduction que si (X_i) est une suite de variables aléatoires i.i.d. de loi de Bernoulli de paramètre p , alors la variable aléatoire

$$Y_n = \frac{S_n - np}{\sqrt{np(1-p)}}, \quad \text{où } S_n = X_1 + \cdots + X_n,$$

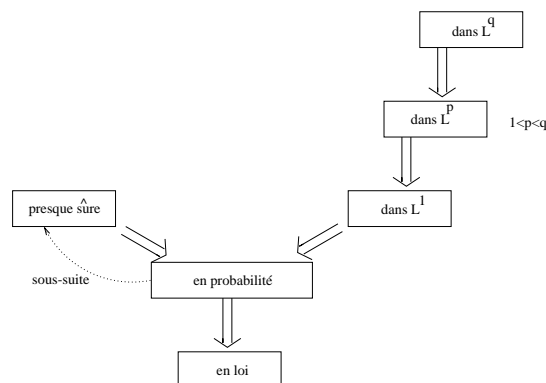
converge en loi vers une gaussienne centrée réduite. Ce résultat est un cas particulier du théorème central limite énoncé au §7. La convergence en loi n'implique pas en général que $P(Y_n \in A)$ converge vers $P(Y \in A)$, où Y est une variable de loi gaussienne centrée réduite. Ceci est vrai si $A = (a, b]$, avec $a \leq -\infty$. Mais considérons l'ensemble $dA = \{(k - np)/\sqrt{np(1-p)} : k, n \in \mathbb{N}\}$; alors pour tout n , $P(Y_n \in A) = 1$, tandis que $P(Y \in A) = 0$. Le problème provient de la frontière ∂A de A , qui dans le dernier cas est \mathbb{R} , tandis que $\partial(a, b] = \{a, b\}$. On montre [2] en fait qu'il y a équivalence entre Y_n converge en loi vers Y et pour toute partie mesurable A telle $P(\partial A) = 0, P(Y_n \in A)$ converge vers $P(Y \in A)$ [2].

Remarque 4.3. Si X_n converge p.s. vers X , alors X_n converge en loi vers X . En effet, d'après le théorème de convergence dominée, pour tout n , la variable aléatoire $Y_n = \exp(itX_n)$ est dominée par la v.a.r. constante égale à 1. Donc par passage à la limite, φ^{X_n} converge vers $\varphi^X(t)$.

De même si X_n converge en probabilité vers X alors X_n converge en loi vers X .

5. Comparaison des convergences

Nous résumons le lien entre les différents types de convergence dans le diagramme suivant.



5.1. Compléments. Nous avons vu que la convergence en probabilité n'implique pas la convergence dans L^1 . Une condition supplémentaire est nécessaire pour avoir équivalence ; celle-ci est la condition d'équi-intégrabilité.

Définition 5.1. Une famille quelconque $(X_i)_{i \in I}$ de variables aléatoires réelles, intégrables et définies sur (Ω, \mathcal{A}, P) est dite *équi-intégrable* ou *uniformément intégrable* si

$$\lim_{c \rightarrow \infty} \sup_{i \in I} \int_{|X_i| > c} |X_i| dP = 0.$$

En particulier, s'il existe une v.a.r. intégrable Y , majorant uniformément la famille (X_i) , c.-à-d. que pour tout $i \in I$, $|X_i| \leq Y$, alors la famille (X_i) est équi-intégrable. En effet,

$$|X_i| \leq Y \implies \int_{|X_i| > c} |X_i| dP \leq \int_{Y > c} Y dP \rightarrow 0.$$

Nous avons la proposition suivante dont la preuve figure par exemple dans [1] ou [11]

Proposition 5.2. Une famille $(X_i)_{i \in I}$ de variables aléatoires réelles, intégrables et définies sur (Ω, \mathcal{A}, P) est équi-intégrable si et seulement si elle vérifie les 2 conditions :

- (i) (*équi-continuité*) pour tout $\epsilon > 0$, il existe un $\eta > 0$ tel que pour tout $A \in \mathcal{A}$, $P(A) \leq \eta$ implique

$$\forall i \in I, \quad \int_A |X_i| dP \leq \epsilon,$$

- (ii) $\sup_{i \in I} \int_{\Omega} |X_i| dP < \infty$.

L'intérêt de la notion d'équi-intégrabilité réside dans le théorème suivant.

Théorème 5.3. Soient $(X_n)_{n \in \mathbb{N}}$, X des variables aléatoires réelles définies sur (Ω, \mathcal{A}, P) . Alors, il y a équivalence entre les deux assertions suivantes :

- (i) $X_n \xrightarrow{P} X$ et la famille (X_n) est équi-intégrable,
(ii) X est intégrable et X_n converge vers X dans L^1 .

Nous en déduisons le corollaire suivant.

Corollaire 5.4. Soit $(X_n)_{n \in \mathbb{N}}$ une suite de variables aléatoires réelles, définies sur (Ω, \mathcal{A}, P) , telles que pour un $p > 1$, $\sup_n E(|X_n|^p) < \infty$. Si $X_n \xrightarrow{P} X$, alors pour tout $q \leq p$, X_n converge vers X dans L^q .

DÉMONSTRATION. Pour tout $c > 0$ et tout $n > 0$

$$\int_{|X_n|^q > c} |X_n|^q dP \leq c^{(q-p)/q} \int_{|X_n|^q > c} |X_n|^p dP \leq c^{(q-p)/q} \sup_{k \in \mathbb{N}} E(|X_k|^p).$$

Par hypothèse et comme $q < p$, le terme de droite tend vers 0 lorsque c tend vers l'infini, uniformément en n . Donc la suite $(|X_n|^q)$ est équi-intégrable et par conséquent, il en est de même de la suite $(|X_n - X|^q)$. Si $X_n \xrightarrow{P} X$, le théorème précédent montre que la suite $(|X_n - X|^q)$ converge dans L^1 vers 0 et donc que la suite $(|X_n - X|)$ converge vers 0 dans L^q . \square

6. Lois faible et forte des grands nombres

Dans toute cette partie, nous désignerons par $(X_n)_{n \geq 1}$ une suite de variables aléatoires réelles indépendantes, de même loi. Pour tout $n \geq 1$, on pose $S_n = \sum_{1 \leq i \leq n} X_i$ et on s'intéressera aux propriétés asymptotiques de S_n . Nous avons déjà vu que la moyenne empirique S_n/n converge en probabilité vers l'espérance $E(X_1)$, lorsque celle-ci existe et sous l'hypothèse de v.a.r. de carré intégrable. Nous avons en fait le théorème suivant.

Théorème 6.1 (Loi faible des grands nombres). *Si $E(|X_1|) < \infty$, alors S_n/n converge en probabilité vers $E(X_1)$ lorsque n tend vers l'infini.*

Avant de prouver ce théorème, démontrons le lemme suivant.

Lemme 6.2. *Si X_n converge en loi vers une variable constante c , alors X_n converge en probabilité vers c .*

DÉMONSTRATION. Pour tout $\epsilon > 0$,

$$\lim_{n \rightarrow \infty} P\{c - \epsilon \leq X_n \leq c + \epsilon\} = \lim_{n \rightarrow \infty} [F^{X_n}(c + \epsilon) - F^{X_n}(c - \epsilon)] = 1$$

et donc $\lim_{n \rightarrow \infty} P\{|X_n - c| \geq \epsilon\} = 0$. □

PREUVE DE LA LOI FAIBLE DES GRANDS NOMBRES. On peut en toute généralité supposer les variables X_n centrées. Puisque X_1 est dans L^1 , la fonction caractéristique est dérivable (cf. proposition 2-(5.3)) et $(\varphi^{X_1})'(0) = E(X_1) = 0$. La formule de Taylor donne $\varphi^{X_1}(t) = 1 + o(t)$, donc par l'indépendance des variables

$$\varphi^{S_n/n}(t) = (\varphi^{X_1}(t/n))^n = (1 + o(n^{-1}))^n = 1 + o(1).$$

Or 1 est la fonction caractéristique de la mesure de Dirac en 0. Donc S_n/n converge en loi vers la constante 0, on conclut grâce au lemme précédent. □

Comme expliqué en introduction à ce chapitre, la loi faible des grands nombres n'est pas satisfaisante. En fait, sous les mêmes hypothèses, le théorème ci-dessus peut être renforcé pour donner la loi forte des grands nombres.

Théorème 6.3 (Loi forte des grands nombres). *Les deux conditions suivantes sont équivalentes :*

- (i) $E(|X_1|) < \infty$,
- (ii) $\lim_{n \rightarrow \infty} S_n/n = E(X_1)$ p.s.

DÉMONSTRATION. Nous ne démontrerons que l'implication (ii) \implies (i), l'autre sens étant plus délicat (cf. [1]). Si la suite S_n/n converge p.s. vers $E(X_1)$, alors X_n/n converge vers 0 p.s. En effet $X_n/n = S_n/n - S_{n-1}/(n-1)$. On en déduit donc que $P\{\lim_n X_n/n = 0\} = 1$. D'après le lemme de Borel-Cantelli 1.2, pour tout $\epsilon > 0$

$$\sum_{n \geq 1} P\{|X_n| \geq \epsilon n\} = \sum_{n \geq 1} P\{|X_n| \geq \epsilon n\} < \infty.$$

On conclut d'après la proposition 2-(4.2).

Esquissons seulement la preuve de l'implication réciproque telle que présentée dans [1]. Dans un premier temps, il suffit de supposer les variables aléatoires centrées. La première étape consistera à supposer que $E(|X_1|^4) < \infty$, puis à approximer toute variable aléatoire dans L^1 par des variables aléatoires dans L^4 (ou même bornées).

Si $E(|X_1|^4) < \infty$ et $E(X_1) = 0$, alors l'inégalité de Markov montre que pour tout $n \geq 1$ et tout $\delta > 0$,

$$P\{|S_n| \geq \delta n\} \leq \frac{1}{\delta^4 n^4} E(S_n^4).$$

L'indépendance et le centrage des X_i conduisent après un calcul simple à

$$E(S_n^4) = nE(X_1^4) + 3n(n-1)(E(X_1^2))^2.$$

Donc $\sum_n P\{|S_n| > n\delta\} < \infty$, ce qui prouve la loi forte des grands nombres grâce au lemme de Borel–Cantelli. Pour la deuxième étape, plus technique, voir [1]. \square

Exemple 61. Donnons une application de la loi forte des grands nombres en théorie des nombres. Considérons l'espace probabilisé $([0, 1], \mathcal{B}([0, 1]), \lambda)$, où λ est la mesure de Lebesgue sur $[0, 1]$. Soit U , la fonction identité, c'est donc une variable aléatoire de loi uniforme sur $[0, 1]$. Pour un $\omega \in [0, 1]$, le développement dyadique du nombre réel $U(\omega)$,

$$U(\omega) = \sum_{i \leq 1} 2^{-i} U_i(\omega),$$

avec $U_i = [2^{i+1}U] - [2^i U]$ où $[\cdot]$ désigne la partie entière (ainsi $U_i = 0$ ou 1). Il y a donc correspondance entre les nombres réels $x \in [0, 1]$ et les suites (u_n) , avec $u_i = 0$ ou 1. Cette correspondance est biunivoque, sauf pour les nombres qui sont de la forme $k/2^n$. Par exemple

$$1/2 = (1, 0, \dots, 0, \dots) = (0, 1, \dots, 1, \dots).$$

Or, l'ensemble des nombres de la forme $k/2^n$, étant dénombrable, est de mesure de Lebesgue nulle; de la même manière l'ensemble des suites (u_n) constantes à partir d'un certain rang est de mesure nulle. On peut donc exclure cet ensemble des suites constantes à partir d'un certain rang.

Les U_i sont mesurables, puisque images de U par des applications mesurables. De plus

$$\begin{aligned} P\{(U_1, \dots, U_n) = (u_1, \dots, u_n)\} &= \lambda\{x \in [0, 1] : (x_1, \dots, x_n) = (u_1, \dots, u_n)\} \\ &= 2^{-n}. \end{aligned}$$

En particulier, les U_i sont des variables indépendantes de loi de Bernoulli de paramètre $1/2$. Notons la correspondance, en cette représentation par un tel "schéma" de Bernoulli et la mesure de Lebesgue sur $[0, 1]$.

La loi forte des grands nombres s'applique, autrement dit presque tout nombre de $[0, 1]$ admet en moyenne autant de 0 et de 1 dans son développement dyadique. Les nombres de $[0, 1]$ possédant la propriété de vérifier la loi forte des grands nombres ont été appelés nombres normaux par Borel.

Comme conclusion, citons E. Borel : « Le jeu de pile ou face, dont le principe est si simple, possède un très grand caractère de généralité et conduit, lorsqu'on l'étudie en détail, aux mathématiques les plus élevées. »

6.1. Extension de la loi des grands nombres. Nous n'avons considéré jusqu'ici que des variables aléatoires identiquement distribuées. Nous supposerons dorénavant que les X_n sont des variables aléatoires mutuellement indépendantes dont chacune possède une espérance et une variance finies

$$\mu_k = E(X_k), \quad \sigma_k^2 = \text{Var}(X_k).$$

La somme S_n a pour espérance m_n et variance s_n^2 données par

$$m_n = \mu_1 + \cdots + \mu_n, \quad s_n^2 = \sigma_1^2 + \cdots + \sigma_n^2.$$

On dira que la suite (X_n) vérifie la loi faible des grands nombres si la suite $(S_n - m_n)/n$ converge en probabilité vers 0. Par application de l'inégalité de Tchebicheff, nous obtenons

$$P\{|S_n - m_n| > n\epsilon\} \leq \frac{1}{n^2\epsilon^2} \text{Var}(S_n - m_n) \leq \frac{s_n^2}{n^2\epsilon^2}.$$

Par conséquent, une condition suffisante pour que la suite (X_n) vérifie la loi faible des grands nombre est que s_n/n converge vers 0 quand n tend vers l'infini. Cette condition n'est pas cependant nécessaire. Par exemple [6], soit la suite (X_n) de variables aléatoires mutuellement indépendantes telles que $X_n = \pm 1$ avec probabilité $(1 - 2^{-n})/2$ et $X_n = \pm 2^n$ avec probabilité 2^{-n-1} . Prouver que (X_n) vérifie les loi faible et forte des grands nombres sans satisfaire la condition précédente.

Pour terminer, signalons que la loi des grands nombres s'applique également dans des conditions plus générales. Ces généralisations, bien que fondamentales (chaînes de Markov, théorie ergodique), sont hors des propos de ce cours.

7. Théorème central limite

Comme précédemment, nous désignerons par $(X_n)_{n \geq 1}$ une suite de variables aléatoires réelles indépendantes et de même loi. La loi forte des grands nombres nous dit que si $E(|X_1|) < \infty$, alors $S_n/n = E(X_1) + o(1)$ p.s. lorsque n tend vers l'infini. Le théorème central limite va préciser le comportement limite en loi du terme $o(1)$.

Théorème 7.1 (central limite). *Sous les hypothèses précédentes, nous avons :*

- (i) si $E(X_1^2) < \infty$, alors $\frac{S_n - nE(X_1)}{\sqrt{n \text{Var}(X_1)}}$ converge en loi vers une variable de loi $\mathcal{N}(0, 1)$;
- (ii) si S_n/\sqrt{n} converge en loi, alors $E(X_1) = 0$, $E(X_1^2) < \infty$ et la loi limite est normale centrée, de variance $\text{Var}(X_1)$.

DÉMONSTRATION. (i) On peut remplacer X_i par $(X_i - E(X_i)) / \sqrt{\text{Var}(X_i)}$, sans perte de généralité, donc supposer que les X_i sont centrées de variance égale à 1. Il suffit alors de montrer que pour tout $t \in \mathbb{R}$,

$$\lim_{n \rightarrow \infty} \varphi^{S_n/\sqrt{n}}(t) = e^{-t^2/2}.$$

Or l'indépendance et l'équidistribution des variables aléatoires X_i donnent

$$\varphi^{S_n/\sqrt{n}}(t) = (\varphi^X(t/\sqrt{n}))^n,$$

où X est une v.a.r. ayant la même loi que les X_i . Par hypothèse, X est de carrée intégrable, donc sa fonction caractéristique est dérivable deux fois et

$$(\varphi^X)'(0) = E(X) = 0, \quad (\varphi^X)''(0) = -E(X^2) = -1.$$

Donc

$$\varphi^X(t) = 1 - \frac{t^2}{2} + o(t^2), \quad \text{lorsque } t \rightarrow 0.$$

Lorsque $n \rightarrow \infty$,

$$\varphi^{S_n/\sqrt{n}}(t) = \left(1 - \frac{t^2}{2} + o(n^{-1})\right)^n = e^{-t^2/2} + o(1).$$

(ii) est admis. □

7.1. Extension du théorème central limite. Le théorème central limite s'étend à des sommes de variables aléatoires indépendantes de loi différentes. En particulier, A. Liapounov a démontré le théorème central limite pour les conditions suivantes :

$$E(X_i) = 0, \quad \text{Var}(X_i) = \sigma_i^2, \quad E(|X_i|^3) = b_i$$

$$\lim_{n \rightarrow \infty} \frac{\sum_{i=1}^n b_i}{\left(\sum_{i=1}^n \sigma_i^2\right)^{3/2}} = 0.$$

La condition la plus générale (nécessaire et suffisante) pour que le théorème central limite soit vrai est la condition de Lindeberg.

Théorème 7.2 (Lindeberg). *Soit (X_i) une suite de v.a.r. mutuellement indépendantes de fonctions de répartition respectives F_1, F_2, \dots telles que*

$$E(X_i) = 0, \quad \text{Var}(X_i) = \sigma_i^2.$$

Posons

$$s_n^2 = \sigma_1^2 + \dots + \sigma_n^2.$$

Supposons que pour tout $t > 0$

$$(7.1) \quad \lim_{n \rightarrow \infty} \frac{1}{s_n^2} \sum_{i=1}^n \int_{|y| \geq ts_n} y^2 F_k(dy) = 0,$$

ou bien la condition suivante équivalente

$$\lim_{n \rightarrow \infty} \frac{1}{s_n^2} \sum_{i=1}^n \int_{|y| < ts_n} y^2 F_k(dy) = 1.$$

Alors, la somme normalisée

$$S_n^* = (X_1 + \cdots + X_n)/s_n$$

converge en loi vers une variable aléatoire gaussienne $\mathcal{N}(0, 1)$.

Pour une preuve, voir [7]. La condition de Lindeberg (7.1) garantit que les variances individuelles σ_i^2 sont petites comparées à leur somme s_n^2 , dans le sens que pour $\epsilon > 0$ fixé et tout n suffisamment grand

$$\sigma_k < \epsilon s_n, \quad k = 1, 2, \dots, n.$$

En effet, σ_k^2/s_n^2 est inférieur à t^2 plus le terme de gauche de 7.1, et en prenant $t = \epsilon/2$ on obtient l'inégalité précédente.

Notons que la condition de Lindeberg est satisfaite si pour un $\delta > 0$, on a pour tout k , $\alpha_k = E(|X_k|^{2+\delta})$ existe et $\alpha_1 + \cdots + \alpha_n = o(s_n^{2+\delta})$. Cette condition est aussi dénommée condition de Liapounov. On retrouve la précédente avec $\delta = 1$. En effet, la somme dans la condition de Lindeberg est majorée par

$$\sum_{i=1}^n \int_{|y| \geq t s_n} \frac{y^2}{t^\delta s_n^{2+\delta}} F_k(dy) \leq \frac{1}{t^\delta} \sum_{i=1}^n \frac{1}{s_n^{2+\delta}} \alpha_k.$$

Comme pour la loi des grands nombres, le théorème central limite s'applique également pour une somme de variables aléatoires, non nécessairement indépendantes (chaîne de Markov, etc.).

8. Compléments

Considérons encore une suite de v.a.r. centrées, indépendantes et identiquement distribuées. On supposera en outre que $E(X^2) = 1$. La loi des grands nombres stipule que S_n/n converge p.s. vers 0, et le théorème central limite nous dit que S_n/\sqrt{n} converge en loi vers $\mathcal{N}(0, 1)$. Que se passe-t-il pour des normalisations de S_n entre $1/n$ et $1/\sqrt{n}$? Cela revient à ce demander avec quelle vitesse a lieu la convergence de S_n/n vers 0. Des estimations de cette vitesse ont été successivement obtenues par Hausdorff (1913), Hardy et Littlewood (1914) et enfin Khintchine (1923) qui obtient le meilleur résultat possible (dans le cas de variables de Bernoulli), connu sous le nom de loi du logarithme itéré. Avant d'énoncer celle-ci, signalons que si a_n/\sqrt{n} tend vers ∞ avec n , alors S_n/a_n tend en probabilité vers 0 (il suffit d'utiliser l'inégalité de Markov). Ceci est encore vrai presque sûrement si $a_n = n^{1/p}$ avec $1 < p < 2$. Mais ce n'est plus le cas si a_n est trop proche de \sqrt{n} . Le cas limite est obtenu pour $a_n = \sqrt{2n \log \log n}$. Le résultat de Khintchine, généralisé par Hartman et Wintner (1941) est le suivant.

Théorème 8.1 (Loi du logarithme itéré). *Soit (X_i) une suite de v.a.r. centrées indépendantes et identiquement distribuées. Supposons que $E(X_i^2) =$*

1. *Alors presque sûrement*

$$\limsup \frac{S_n}{\sqrt{2n \log \log n}} = 1$$

$$\liminf \frac{S_n}{\sqrt{2n \log \log n}} = -1$$

En d'autres termes, pour tout $\epsilon > 0$, la suite $S_n(\omega)$ ne dépasse jamais à partir d'un certain rang les niveaux extrêmes $\pm 2n(1 + \epsilon) \log \log n$, mais par contre atteint une infinité de fois $\pm 2n(1 - \epsilon) \log \log n$ ($0 < \epsilon < 1$), pour tout les ω n'appartenant pas à un ensemble de probabilité nulle.

Cette loi met en évidence les fluctuations des sommes partielles S_n de la suite, qui traversent, avec probabilité un, une infinité de fois le niveau 0, tout en prenant successivement des valeurs positives ou négatives arbitrairement grandes (on peut cependant remarquer le faible taux de croissance de la fonction $\log \log x$).

Comme autre implication de la loi du logarithme itéré, supposons qu'il existe une v.a.r. Z de loi $\mathcal{N}(0, 1)$, telle que la v.a.r. $Z_n = S_n/\sqrt{n}$ converge en probabilité vers Z . Quitte à extraire une sous-suite, nous pouvons supposer que Z_n converge p.s. vers Z . Or la loi du logarithme itéré contredit le fait que $\lim_{n \rightarrow \infty} Z/\sqrt{2 \log \log n} = 0$. Ainsi, la convergence en loi ne concerne que les lois et non les variables.

Il existe d'autres loi limites, en particulier pour les valeurs extrêmes, c.-à-d. pour les v.a.r.

$$m_n = \inf\{X_1, \dots, X_n\}, \quad M_n = \sup\{X_1, \dots, X_n\}.$$

En effet, un théorème de Glivenko (1943) prouve que les extrêmes convergent asymptotiquement pour pratiquement toutes les distributions connues (à l'exception des distributions discrètes) vers des lois limites qui sont (pour le maximum M_n) :

– la loi de Fréchet :

$$P(M < x) = \exp(-x^{-a}), \quad x > 0, a > 0;$$

– la loi de Weibull :

$$P(M < x) = 1 - \exp(-(-x)^{-a}), \quad x < 0, a > 0;$$

– la loi de Gumbel :

$$P(M < x) = \exp(-e^{-x}), \quad -\infty < x < \infty.$$

Dans le cas particulier [5], où les X_i sont indépendantes de loi $\mathcal{N}(0, 1)$, alors on obtient plus précisément :

$$(8.1) \quad \lim_{n \rightarrow \infty} P\left(\sqrt{2 \log n}(M_n - \sqrt{2 \log n} + \frac{\log \log n + \log 4\pi}{2\sqrt{2 \log n}}) < x\right) = e^{-e^{-x}}.$$

La suite M_n se concentre donc sur

$$\sqrt{2 \log n} + \frac{\log \log n + \log 4\pi}{2\sqrt{2 \log n}}.$$

Il est en fait possible de prouver qu'avec probabilité un, à partir d'un certain rang, on a les inégalités :

$$\sqrt{2 \log n} - \frac{\log \log n}{2\sqrt{2 \log n}}(1 + \epsilon) < M_n < \sqrt{2 \log n} + \frac{\log \log n}{2\sqrt{2 \log n}}(1 + \epsilon),$$

pour un ϵ fixé, les encadrements étant les meilleures possibles (ils ne seraient pas satisfaits pour $\epsilon = 0$). Ceci implique en particulier, que $\lim(M_n - \sqrt{2 \log n}) = 0$ p.s.

Bibliographie

- [1] Ph. Barbe and M. Ledoux. *Probabilité*. Collection mathématiques : de la licence à l'agrégation. Editions Belin et Editions Espaces 34, 1998.
- [2] P. Billingsley. *Convergence of probability measures*. Wiley, N.Y., 1968.
- [3] Nicolas Bouleau. *Probabilités de l'ingénieur*. Hermann, 1986.
- [4] C. Coccozza-Thivent. *Processus stochastiques et fiabilité des systèmes*. SMAI. Springer, 1997.
- [5] P. Deheuvels. *La probabilité le hasard et la certitude*, volume 3 of *Que sais-je ?* P.U.F., 1990.
- [6] W. Feller. *An Introduction to Probability Theory and Its Applications*, volume 1. Wiley & Sons, 3ème édition, 1971.
- [7] W. Feller. *An Introduction to Probability Theory and Its Applications*, volume 2. Wiley & Sons, 2ème édition, 1971.
- [8] M. Kac. *Probability and related topics in physical sciences*. Interscience publishers, LTD., 1957.
- [9] D. E. Knuth. *The Art of Computer programming*, volume 2. Addison-Wesley, 1981.
- [10] B. Lacaze, C. Mailhes, M. Maubourguet, and J.-Y. Tourneret. *Probabilités et Statistique Appliquées*. CÉPADUÈS-ÉDITIONS, 1997.
- [11] J. Neveu. *Bases mathématiques du calcul des probabilités*. Masson, 2ème édition, 1980.
- [12] Russell A. Paielli and H Erzberger. Conflict Probability Estimation For Free Flight. *AIAA Journal of Guidance, Control and Dynamics*, 1996.
- [13] Pour la Science, Dossier hors-série. *Le hasard*, 1996.
- [14] H. Ventsel. *Théorie des probabilités*. Mir, 1973.