PROBABILITES

Jean JACOD

1999-2000

CHAPITRE 1

Introduction aux probabilités

1 Rappels sur les ensembles

Considérons un ensemble E, c'est-à-dire une collection d'objets appelés les "éléments", ou les "points", de E. L'appartenance d'un point x à l'ensemble E est notée $x \in E$, et $x \notin E$ signifie que le point x n'appartient pas à E.

Une partie de E est aussi un ensemble, appelé sous-ensemble de E: on écrit $F \subset E$ (on dit aussi que F est "inclus" dans E) lorsque F est un sous-ensemble de E.

Rappelons les opérations élémentaires sur les parties d'un ensemble:

Intersection: $A \cap B$ est l'intersection des ensembles A et B, i.e. l'ensemble des points appartenant à la fois à A et à B.

Réunion: $A \cup B$ est la réunion des ensembles A et B, i.e. l'ensemble des points appartenant à au moins l'un de ces deux ensembles.

Complémentaire: Si $A \subset E$, son complémentaire (dans E) est l'ensemble des points de E n'appartenant pas à A; on le note A^c, ou parfois $E \backslash A$.

Différence: Si A et B sont deux sous-ensembles de E, tels que $A \subset B$, on note $A \backslash B$ la "différence" entre A et B, i.e. l'ensemble des points qui sont dans A mais pas dans B; on a donc $A \backslash B = A \cap B^c$.

Différence symétrique: $A \Delta B$ est l'ensemble des points appartenant à l'un des deux ensembles A ou B, mais pas aux deux; on a donc $A \Delta B = (A \backslash (A \cap B)) \cup (B \backslash (A \cap B))$.

Ensemble vide: C'est l'ensemble ne contenant aucun point; on le note $\emptyset$.

Ensembles disjoints: Les ensembles A et B sont dits *disjoints* si $A \cap B = \emptyset$.

La réunion et l'intersection sont des opérations commutatives et associatives: on a $A \cup B = B \cup A$ et $A \cap B = B \cap A$, et aussi $A \cup (B \cup C) = (A \cup B) \cup C$ et $A \cap (B \cap C) = (A \cap B) \cap C$, ensembles qu'on note naturellement $A \cup B \cup C$ et $A \cap B \cap C$. Plus généralement si on a une famille $(A_i)_{i \in I}$ d'ensembles, indexée par un ensemble quelconque I, on note $\cup_{i \in I} A_i$

(resp. $\cap_{i\in I}A_i$) la réunion (resp. l'intersection) de cette famille, i.e. l'ensemble des points appartenant à au moins l'un des A_i (resp. appartenant à tous les A_i): l'ordre d'indexation des A_i n'a pas d'importance.

Les ensembles suivants seront utilisés sans cesse:

$\mathbb{N}$ = ensemble des entiers naturels: $0, 1, 2, ...$

$\mathbb{Z}$ = ensemble des entiers relatifs: $..., -2, -1, 0, 1, 2, ...$

$\mathbb{Q}$ = ensemble des rationnels

$\mathbb{R}$ = ensemble des réels $=]-\infty, \infty[$

$\mathbb{R}^d$ = espace euclidien réel de dimension d (donc $\mathbb{R}^1 = \mathbb{R}$)

$\overline{\mathbb{R}} = [-\infty, \infty]$

$\mathbb{R}_+ = [0, \infty[$

$\overline{\mathbb{R}}_+ = [0, \infty]$

$\mathbb{N}^*$ = ensemble des entiers naturels non nuls: $1, 2, ...$

$\mathbb{C}$ = ensemble des nombres complexes.

L'ensemble des points a_i indexés par un ensemble I est noté $\{a_i : i \in I\}$. Si on a un nombre fini de points $a_1, ..., a_n$, on écrit aussi $\{a_1, a_2, ..., a_n\}$. Par exemple, on a $\mathbb{N}^* = \mathbb{N}\backslash\{0\}$.

2 Phénomènes aléatoires

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

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

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

Exemples:

1) Un tirage à pile ou face: $\Omega = \{\text{p,f}\}$.

2) Deux tirages successifs à pile ou face: $\Omega = \{\text{pp,pf,fp,ff}\}$.

3) Tirage de deux dés: $\Omega = \{(i,j) : 1 \leq i \leq 6, 1 \leq j \leq 6\}$.

4) Mesure d'une longueur L, avec une erreur de mesure: $\Omega = \mathbb{R}_+$; $\omega \in \Omega$ désigne le résultat de la mesure, et $\omega - L$ est l'erreur de mesure.

5) Durée de vie d'une ampoule électrique: $\Omega = \mathbb{R}_+$.

b) Les événements: Un événement est une propriété dont on peut dire si elle est vraie ou non, une fois l'expérience réalisée. En termes mathématiques, un événement est une partie de Ω. Si A et B sont deux événements,

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

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

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

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

$$P(A) \;=\; \text{limite de } \; f_n(A) \; \text{ quand } \; n \uparrow +\infty. \tag{1}$$

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

(P0) $\quad 0 \leq P(A) \leq 1$,

(P1) $\quad P(\Omega) = 1$,

(P2) $\quad P(A \cup B) = P(A) + P(B) \quad$ si $\quad A \cap B = \emptyset$,

et il en découle:

$$P(\emptyset) \;=\; 0 \quad \text{(appliquer (P2) avec } \; A = B = \emptyset), \tag{2}$$

$$P(A) + P(A^c) = 1, \tag{3}$$

$$P(\cup_{i=1}^n A_i) = \sum_{i=1}^n P(A_i) \quad \text{si les } A_i \text{ sont deux-à-deux disjoints,} \tag{4}$$

$$P(A \cup B) + P(A \cap B) = P(A) + P(B), \tag{5}$$

$$P(A) \leq P(B) \quad \text{si } A \subset B. \tag{6}$$

Un modèle probabiliste est donc un triplet $(\Omega, \mathcal{A}, P)$, constitué de l'espace Ω, de l'ensemble des événements $\mathcal{A}$, et de la famille des $P(A)$ pour $A \in \mathcal{A}$: on peut ainsi considérer P comme une application de $\mathcal{A}$ dans $[0,1]$, qui vérifie au moins les propriétés (P1) et (P2) ci-dessus (plus une propriété supplémentaire, plus difficile à comprendre, et qui sera expliquée plus bas).

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

d) Variable aléatoire: il s'agit là d'une grandeur qui dépend du résultat de l'expérience. En termes mathématiques, c'est une application de Ω dans un espace E, en général $E = \mathbb{R}$ ou $E = \mathbb{R}^d$. **Attention:** cette terminologie, consacrée par l'usage, est très malencontreuse; une "variable" aléatoire n'et pas une variable (au sens de l'analyse), mais une fonction !

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

$$P_X(B) = P(X^{-1}(B)) \quad \text{pour} \quad B \subset E, \tag{7}$$

où $X^{-1}(B)$ désigne l'image réciproque de B par X, c'est-à-dire l'ensemble des $\omega \in \Omega$ tels que $X(\omega) \in B$. Cette formule définit une nouvelle probabilité, notée P_X, cette fois-ci sur l'espace E (au lieu de Ω). Cette probabilité P_X s'appelle le **loi de la variable** X.

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

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

où $\text{card}(A)$ désigne le nombre de points contenus dans A. On vérifie aisément (P0), (P1) et (P2), et on a $P(\{\omega\}) = \frac{1}{36}$ pour chaque singleton. L'application $X : \Omega \to \mathbb{N}$ définie par $X(i,j) = i + j$ est la variable aléatoire "somme des deux dés", de loi

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

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

3 Espace d'états fini et combinatoire

Dans ce paragraphe on suppose que l'espace d'états Ω est **fini**. On prend alors $\mathcal{A} = \mathcal{P}(\Omega)$.

Définition 1: Une **probabilité** sur Ω fini est une application $P : \mathcal{P}(\Omega) \to [0,1]$ qui vérifie (P1) et (P2). On a donc également les propriétés (2)-(6).

Proposition 2: *a) Une probabilité P sur Ω fini est entièrement caractérisée par ses valeurs sur les singletons, soit $p_\omega = P(\{\omega\})$.*

b) Etant donnée une famille $(p_\omega)_{\omega\in\Omega}$ de réels, il lui correspond une probabilité P (nécessairement unique) telle que $P(\{\omega\}) = p_\omega$ pour tout $\omega \in \Omega$ si et seulement si

$$p_\omega \geq 0, \qquad \sum_{\omega\in\Omega} p_\omega = 1. \tag{8}$$

et dans ce cas, on a pour tout $A \in \Omega$:

$$P(A) = \sum_{\omega\in A} p_\omega \qquad \text{(avec la convention} \quad \sum_{\emptyset} = 0\ ). \tag{9}$$

Preuve. Soit d'abord P une probabilité sur Ω, et soit $p_\omega = P(\{\omega\})$. Il est alors évident que $p_\omega \geq 0$, et (9) découle de (4), puisque n'importe quelle partie A de Ω est réunion disjointe des singletons $\{\omega\}$, pour les $\omega \in A$. On a donc (a) et la condition nécessaire de (b) (la seconde partie de (8) découle de (9) appliqué à $A = \Omega$ et de $P(\Omega) = 1$).

Inversement, soit des p_ω vérifiant (8). On définit $P(A)$ pour tout $A \subset \Omega$ par (9), et la vérification de (P0), (P1) et (P2) est immédiate. $\square$

Définition 3: On dit que la probabilité P sur l'espace fini Ω est **uniforme** si $p_\omega = P(\{\omega\})$ ne dépend pas de ω.

Si P est uniforme,il découle de (8) et (9) que

$$P(A) = \frac{\text{card}(A)}{\text{card}(\Omega)}, \tag{10}$$

de sorte que le calcul des probabilités se ramène à des dénombrements: on est dans le cas de la **combinatoire**. Remarquer d'ailleurs que sur un espace fini donné Ω il existe une et une seule probabilité uniforme.

Nous allons maintenant donner deux exemples très importants pour les applications.

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

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

Ensuite, il est également naturel de considérer que tous les tirages possibles sont équiprobables, donc P est la probabilité uniforme sur Ω. La quantité X est une variable aléatoire car si on connait le tirage ω, on sait aussi le nombre $X(\omega)$ de boules blanches qu'il contient. L'ensemble $X^{-1}(\{x\})$, noté aussi $\{X = x\}$, contient $C_N^x C_M^{n-x}$ éléments si $x \leq N$ et $n - x \leq M$, et est vide sinon. Donc

$$P(X = x) \;=\; \begin{cases} \frac{C_N^x \, C_M^{n-x}}{C_{N+M}^n} & \text{si } \; 0 \leq x \leq N \;\text{ et }\; 0 \leq n - x \leq M \\ 0 & \text{sinon.} \end{cases} \tag{11}$$

On a ainsi obtenu, lorsque x varie, la **loi** de X. Cette loi, appelée loi hypergéométrique, intervient naturellement dans la théorie des sondages: il y a $N + M$ électeurs, dont N ont l'opinion "blanche" et M l'opinion "noire", et on sonde avec un échantillon de taille n (bien-sûr, il y a en général plus de deux opinions possibles; **exercice:** l'urne contient N_j boules de couleur j, pour $j = 1, \ldots, p$, et on tire n boules, dont X_j seront de couleur j; calculer la probabilité $P(X_1 = x_1, X_2 = x_2, \ldots, X_p = x_p)$).

On voit dès cet exemple l'intérêt de la notion de loi pour une variable aléatoire. Les expériences consistant à tirer des boules dans une urne et à faire un sondage sont très différentes, mais la variable X (nombre de boules blanches dans un cas, d'opinions "blanches" dans l'autre) a même loi dans les deux cas: on peut donc faire abstraction des expériences proprement dites, et raisonner sur la variable X, qui a toujours les mêmes propriétés probabilistes.

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

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

$$P(X = x) \;=\; C_n^x \left(\frac{N}{N+M}\right)^x \left(\frac{M}{N+M}\right)^{n-x} \quad \text{pour} \quad 0 \leq x \leq n. \tag{12}$$

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

$$P(X = x) \;=\; C_n^x \, p^x (1 - p)^{n-x} \quad \text{pour} \quad 0 \leq x \leq n. \tag{13}$$

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

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

$$P(X = x) \;\to\; C_n^x \, p^x (1 - p)^{n-x} \qquad \text{pour} \quad x = 0, 1, \ldots, n, \tag{14}$$

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

4 Définition générale des probabilités

Lorsque l'espace d'états Ω n'est pas fini, la définition 1 est insuffisante.

Considérons un exemple. Si on tire n fois à pile ou face, l'espace Ω naturel est $\Omega = \{\text{p,f}\}^n$ (produit cartésien de n fois l'ensemble $\{\text{p,f}\}$; c'est aussi l'ensembles des "mots" de n lettres, avec un alphabet constitué des deux lettres "p" et "f"). C'est un ensemble fini, avec 2^n éléments. Si le jeu n'est pas truqué, on est dans le cadre de la combinatoire (probabilité uniforme), de sorte que d'après (10) on a

$$P(A) \;=\; \frac{\text{card}(A)}{2^n} \qquad \text{pour} \quad A \subset \Omega. \tag{15}$$

Supposons maintenant qu'on poursuive le jeu indéfiniment. L'espace d'état devient $\Omega = \{\text{p,f}\}^{I\!N^*}$, c'est-à-dire l'ensemble des mots de longueur infinie, avec le même alphabet "p" et "f". C'est un ensemble infini. Essayons alors d'évaluer la probabilité $P(A)$ de l'événement $A =$ "on ne tire jamais pile". Soit $A_n =$ "on ne tire jamais pile lors des n premiers tirages"; d'après (15), on a $P(A_n) = 2^{-n}$. Maintenant, A est la "limite" des A_n, au sens où les A_n sont décroissants (i.e. $A_{n+1} \subset A_n$) et où $A = \cap_n A_n$. Il est alors naturel d'écrire que

$$P(A) \;=\; \lim_{n\uparrow\infty} P(A_n) \;=\; 0. \tag{16}$$

Pour que ceci soit vrai, les propriétés (P0)-(P2) sont insuffisantes; il faut ajouter un axiome supplémentaire permettant le "passage à la limite" dans (16).

A cet effet, il nous faut d'abord caractériser les propriétés que doit satisfaire la classe $\mathcal{A}$ des événements. En effet, si sur un ensemble fini il est naturel de prendre $\mathcal{A} = \mathcal{P}(\Omega)$, il n'en est plus de même lorsque Ω est infini: ceci pour des raisons mathématiques qui seront explicitées plus loin, et aussi pour des raisons de modélisation qui seront également examinées plus bas. Il se trouve que, à l'instar de la probabilité, la classe $\mathcal{A}$ doit satisfaire un certain système d'axiomes. Avant de définir ces axiomes, nous allons d'abord rappeler ce qu'est un ensemble dénombrable:

Les ensembles dénombrables: on dit qu'un ensemble E est *dénombrable* s'il est en bijection avec $I\!N$, c'est-à-dire si on peut énumérer ses points en une suite $(x_n)_{n\in I\!N}$: c'est le cas de $I\!N$ lui-même, ou de $I\!N^*$, de $\mathbb{Z}$, de $\mathbb{Q}$, ou encore des entiers pairs, ou de toute suite strictement croissante d'entiers. Ce n'est pas le cas de $I\!R$, ni des intervalles $[a,b]$ lorsque $a < b$.

Voici quelques propriétés des ensembles dénombrables: d'abord, toute partie d'un ensemble dénombrable est elle-même finie ou dénombrable. La réunion d'une famille finie ou dénombrable d'ensembles eux-mêmes finis ou dénombrables est un ensemble fini ou dénombrable. En revanche si A est n'est ni fini ni dénombrable, il en est de même de $A\backslash B$ pour tout $B \subset A$ qui est fini ou dénombrable. $\square$

Voici maintenant une liste d'axiomes:

(A1) On a $\emptyset \in \mathcal{A}$ et $\Omega \in \mathcal{A}$.

(A2) $\mathcal{A}$ est stable par complémentation, i.e. $A \in \mathcal{A} \Rightarrow A^c \in \mathcal{A}$.

(A3) $\mathcal{A}$ est stable par réunion et intersection finie, i.e. si $A, B \in \mathcal{A}$ alors $A \cup B \in \mathcal{A}$ et $A \cap B \in \mathcal{A}$.

(A4) $\mathcal{A}$ est stable par réunion et intersection dénombrable, i.e. si $(A_n)_{n\in\mathbb{N}}$ est une suite d'éléments de $\mathcal{A}$, alors $\cup_{n\in\mathbb{N}} A_n$ et $\cap_{n\in\mathbb{N}} A_n$ sont dans $\mathcal{A}$.

Bien entendu, ces axiomes sont redondants: par exemple sous (A2), on a (A1) si et seulement si $\emptyset \in \mathcal{A}$, et on a (A3) (resp. (A4)) si et seulement si $\mathcal{A}$ est stable par réunion (ou intersection) finie (resp. dénombrable. Noter que (A4)$\Rightarrow$(A3): il suffit de prendre $A_0 = A$ et $A_n = B$ pour tout $n \geq 1$. Noter aussi que (A4) *n'entraine pas* que $\mathcal{A}$ soit stable par réunion ou intersection infinie non dénombrable.

Définition 4: La classe $\mathcal{A}$ est une **algèbre** si elle vérifie (A1), (A2) et (A3). C'est une **tribu** (ou: σ-algèbre) si elle vérifie (A1), (A2) et (A4) (donc aussi (A3)).

Si l'espace Ω est fini, toute algèbre est une tribu. Cela est faux dès que Ω est infini.

Définition 5: Si $\mathcal{C} \subset \mathcal{P}(\Omega)$, on appelle **tribu engendrée par** $\mathcal{C}$, et on note $\sigma(\mathcal{C})$, la plus petite tribu contenant $\mathcal{C}$: elle existe toujours, car d'une part $\mathcal{P}(\Omega)$ est une tribu, d'autre part l'intersection d'une famille quelconque de tribus est une tribu; ainsi $\sigma(\mathcal{C})$ est l'intersection de toutes les tribus contenant $\mathcal{C}$.

Exemples:

1) $\mathcal{A} = \{\emptyset, \Omega\}$ est la tribu grossière, ou triviale (c'est la plus petite tribu de Ω).

2) La tribu engendrée par l'ensemble $\{A\}$ est $\{\emptyset, A, A^c, \Omega\}$.

3) Si $(A_i)_{i\in I}$ est une partition finie ou dénombrable de Ω (i.e. les A_i sont deux-à-deux disjoints et leur réunion est Ω), la tribu engendrée par $\{A_i : i \in I\}$ est l'ensemble des réunions $B_J = \cup_{i\in J} A_i$, où J décrit la classe de toutes les parties de I (avec la convention que $B_\emptyset = \emptyset$).

4) Si $\Omega = \mathbb{R}$ (ou plus généralement si Ω est un espace topologique), on appelle *tribu borélienne* la tribu engendrée par la classe des ouverts (ou de manière équivalente par la classe des fermés). A titre d'exercice de maniement des tribus, nous donnons en détail la démonstration du résultat suivant:

Proposition 6: *La tribu borélienne de* $\mathbb{R}$ *est aussi la tribu engendrée par les intervalles de la forme* $]-\infty, a]$, *pour* $a \in \mathbb{Q}$.

Preuve. Soit $\mathcal{C}_1$ la classe des intervalles ouverts de $\mathbb{R}$, et $\mathcal{C}_2$ la classe des intervalles $]-\infty, a]$ pour $a \in \mathbb{Q}$. Soit enfin $\mathcal{R}$ la tribu borélienne.

Tout ouvert A est réunion dénombrable d'intervalles ouverts: en effet, on a $A = \cup_{(q,n)\in B}]q - \frac{1}{n}, q + \frac{1}{n}[$, où B est l'ensemble (dénombrable) des couples (q, n) avec $q \in \mathbb{Q}$, $n \in \mathbb{N}^*$ et $]q - \frac{1}{n}, q + \frac{1}{n}[\subset A$. On a donc $\sigma(\mathcal{C}_1) = \mathcal{R}$ par (A4) et le fait que $\mathcal{C}_1 \subset \mathcal{R}$.

Par ailleurs soit $]x, y[\in \mathcal{C}_1$, et soit (x_n) (resp. (y_n)) une suite de rationnels décroissant vers x (resp. croissant strictement vers y). On a

$$]x,y[\ = \ \cup_n \left( ]-\infty, y_n] \cap]-\infty, x_n]^c \right),$$

donc $\mathcal{C}_1 \subset \sigma(\mathcal{C}_2)$, donc $\mathcal{R} = \sigma(\mathcal{C}_1) \subset \sigma(\mathcal{C}_2)$. Comme tout fermé est dans $\mathcal{R}$, on a aussi $\mathcal{C}_2 \subset \mathcal{R}$, donc finalement $\sigma(\mathcal{C}_2) = \mathcal{R}$. $\square$

Passons maintenant à la définition générale des probabilités:

Définition 7: Une **probabilité** sur la tribu $\mathcal{A}$ de Ω est une application $P : \mathcal{A} \to [0,1]$ vérifiant les deux axiomes suivants:

(P1) $P(\Omega) = 1$.

(P3) Pour toute **suite** (dénombrable) (A_n) d'éléments de $\mathcal{A}$ qui sont deux-à-deux disjoints, on a $P(\cup_n A_n) = \sum_n P(A_n)$.

L'axiome (P3), dit "axiome de σ-additivité", est plus fort que (P2), appelé aussi axiome d'additivité: pour le voir, on commence par appliquer (P3) avec $A_n = \emptyset$ pour tout $n \in I\!N$; si $a = P(\emptyset)$ on obtient $a = \sum_{n \in I\!N} a$, ce qui entraine $a = 0$; ensuite, si $A, B \in \mathcal{A}$ sont disjoints, on applique (P3) avec $A_0 = A$, $A_1 = B$ et $A_n = \emptyset$ pour tout $n \geq 2$, ce qui donne $P(A \cup B) = P(A) + P(B) + \sum_{n \geq 2} P(\emptyset) = P(A) + P(B)$, d'où (P2).

Noter que toute probabilité P vérifie les propriétés (2)-(6). Le résultat suivant illustre les rapports entre (P2) et (P3). Pour ce résultat, on utilise les notations suivantes: si (A_n) est une suite de parties de Ω, on écrit $A_n \downarrow A$ (resp. $A_n \uparrow A$) si la suite (A_n) est décroissante, i.e. $A_{n+1} \subset A_n$ pour tout n et si $A = \cap_n A_n$ (resp. est croissante, i.e. $A_n \subset A_{n+1}$ pour tout n et si $A = \cup_n A_n$).

Proposition 8: *Supposons que l'application $P : \mathcal{A} \to [0,1]$ vérifie (P1) et (P2). Les conditions suivantes sont alors équivalentes:*

(i) On a (P3).

(ii) On a $P(A_n) \downarrow 0$ si $A_n \downarrow \emptyset$ et $A_n \in \mathcal{A}$.

(iii) On a $P(A_n) \downarrow P(A)$ si $A_n \downarrow A$ et $A_n \in \mathcal{A}$ (donc $A \in \mathcal{A}$).

(iv) On a $P(A_n) \uparrow 1$ si $A_n \uparrow \Omega$ et $A_n \in \mathcal{A}$.

(v) On a $P(A_n) \uparrow P(A)$ si $A_n \uparrow A$ et $A_n \in \mathcal{A}$ (donc $A \in \mathcal{A}$).

Preuve. Etant donné (3), on a (ii)$\Leftrightarrow$(iv) et (iii)$\Leftrightarrow$(v). On a aussi évidemment (v)$\Rightarrow$(iv). Inversement, supposons (iv), et soit $A_n \in \mathcal{A}$ avec $A_n \uparrow A$. Soit $B_n = A_n \cup A^c$, qui croît vers Ω, donc $P(B_n) = P(A_n) + P(A^c)$ (à cause de (P2), puisque $A_n \cap A^c = \emptyset$) croît vers $1 = P(A) + P(A^c)$: on a donc (v).

Il nous reste à montrer que (i) est équivalent à (v). Supposons d'abord (v). Considérons une suite (A_n) d'éléments de $\mathcal{A}$, deux-à-deux disjoints, et posons $B_n = \cup_{p \leq n} A_p$ et $B = \cup_n A_n$. D'après (4) on a $P(B_n) = \sum_{p \leq n} P(A_p)$, qui croît vers $\sum_n P(A_n)$, et aussi vers $P(B)$ par (v): on a donc (i).

Supposons enfin (i). Soit $A_n \in \mathcal{A}$ avec $A_n \uparrow A$, pour $n \geq 0$. Soit aussi $B_0 = A_0$, et définissons par récurrence $B_n = A_n \backslash B_{n-1}$ pour $n \geq 1$. Comme $\cup_n B_n = A$ et comme les B_n sont deux-à-deux disjoints, on a

$$P(A) = \sum_n P(B_n) = \lim_n \sum_{p=0}^{n} P(B_p) = \lim_n P(A_n),$$

la dernière égalité provenant de (4): on a donc (v). □

La propriété (P3) donne la probabilité de la réunion $\cup_n A_n$ en fonctions des $P(A_n)$, lorsque les A_n sont deux-à-deux disjoints. Si ce n'est pas le cas, on a tout de même le majoration suivante, très utile dans la pratique:

Proposition 9: *Soit P une probabilité, et soit $(A_n)_{n\in I}$ une famille finie ou dénombrable d'événements. On a alors*

$$P(\cup_{n\in I} A_n) \ \leq \ \sum_{n\in I} P(A_n). \tag{17}$$

Preuve. a) Supposons d'abord l'ensemble I fini. Il s'agit de montrer que pour tout k entier on a

$$P(A_1 \cup \ldots \cup A_k) \ \leq \ P(A_1) + \ldots + P(A_k). \tag{18}$$

Nous montrons cette propriété par récurrence sur k: elle est évidente pour $k = 1$. Supposons la vraie pour $k-1$, avec $k \leq 2$, et posons $B = A_1 \cup \ldots A_{k-1}$ et $C = B \cup A_k$. En vertu de (5), on a $P(C) + P(B \cap A_k) = P(B) + P(A_k)$, donc $P(C) \leq P(B) + P(A_k)$, et on en déduit immédiatement que (18) est satisfait pour k.

b) Passons maintenant au cas où I est dénombrable: on peut supposer sans restriction qu'alors $I = I\!N^*$. On pose $B_n = \cup_{i=1}^n A_i$, qui croît vers l'ensemble $C = \cup_{n\in I} A_n$. D'après (a) on a

$$P(B_n) \ \leq \ \sum_{i=1}^{n} P(A_i).$$

Mais le membre de gauche ci-dessus croît vers $P(C)$ en vertu de la proposition précédente, tandis que le membre de droite croît vers $\sum_{n\in I} P(A_n)$: en passant à la limite, on obtient donc (17). □

Pour terminer ce paragraphe, revenons sur les lois des variables aléatoires. Soit l'espace d'états Ω muni de la tribu $\mathcal{A}$ et de la probabilité P. Soit X une application de Ω dans un autre ensemble E (comme on l'a dit, E égale en général $I\!R$ ou $I\!R^d$, ou un sous-ensemble de ces ensembles, tel $I\!N$). Comme $P(A)$ n'est définie que pour les A dans $\mathcal{A}$, la formule (7) ne permet de définir $P_X(B)$ que pour les B tels que $X^{-1}(B) \in \mathcal{A}$, d'où l'intérêt du résultat suivant:

Proposition 10: *a) La famille $\mathcal{E}$ des parties B de E telles que $X^{-1}(B) \in \mathcal{A}$ est une tribu de E.*

b) La formule (7), pour $B \in \mathcal{E}$, définit une probabilité sur la tribu $\mathcal{E}$ de E, appelée la **loi** *de X.*

Preuve. Les propriétés (A1), (A2) et (A3) pour $\mathcal{E}$, ainsi que (P1) et (P3) pour P_X, découlent immédiatement des propriétés de même nom pour $\mathcal{A}$ et P, une fois remarquées les propriétés élémentaires suivantes: $X^{-1}(\emptyset) = \emptyset$, $X^{-1}(E) = \Omega$, $X^{-1}(B^c) = (X^{-1}(B))^c$, $X^{-1}(\cap_i A_i) = \cap_i X^{-1}(A_i)$ et $X^{-1}(\cup_i A_i) = \cup_i X^{-1}(A_i)$. $\square$

On remarquera qu'en général $\mathcal{E} \neq \mathcal{P}(E)$, même si on a $\mathcal{A} = \mathcal{P}(\Omega)$. Comme P_X ne saurait être définie que sur $\mathcal{E}$, ceci constitue une première raison, d'ordre mathématique, pour qu'en général une probabilité soit définie sur une tribu qui peut être strictement plus petite que l'ensemble de toutes les parties.

5 Indépendance et conditionnement

Commençons par l'indépendance. Intuitivement, deux événements A et B sont indépendants si le fait de savoir que A est réalisé ne donne aucune information sur la réalisation de B.

Considérons l'approche par les fréquences: on répète n fois la même expérience. Si A et B sont indépendants, la fréquence $f_n(B)$ de réalisation de B (qui approche $P(B)$ pour n grand) doit être approximativement la même que la fréquence $f_n(B/A)$ de réalisation de B parmi les expériences pour lesquelles A est réalisé. Si on remarque que $f_n(B/A) = \frac{f_n(A\cap B)}{f_n(A)}$ (à condition bien-sûr que $f_n(A) > 0$), en "passant à la limite" en n, on est conduit naturellement à la définition suivante:

Définition 11. a) Deux événements A et B sont **indépendants** si

$$P(A \cap B) \;=\; P(A)P(B). \tag{19}$$

b) Les événements $(A_i)_{i\in I}$ (où I est un ensemble quelconque) sont dits **indépendants** (on dit aussi "mutuellement indépendants") si, pour toute partie *finie* J de I on a

$$P(\cap_{i\in J} A_i) \;=\; \prod_{i\in J} P(A_i). \tag{20}$$

Des événements indépendants sont aussi deux-à-deux indépendants, mais la réciproque peut être fausse. Nous laissons en exercice (très facile) la démonstration de la proposition suivante, dont le résultat est tout-à-fait intuitif:

Proposition 12: *Si A et B sont indépendants, il en est de même des couples (A,B^c), (A^c,B) et (A^c,B^c).*

Exemples.

1) On tire 3 fois un dé. Si A_i est un événement qui ne dépend que du $i^{\text{ème}}$ tirage, alors A_1, A_2 et A_3 sont indépendants.

2) On tire une carte au hasard dans un jeu de 52 cartes. Soit $A =$ "la carte tirée est un pique" et $B =$ "la carte tirée est un as". Alors A et B sont indépendants.

3) Soit $\Omega = \{1,2,3,4\}$ avec $\mathcal{A} = \mathcal{P}(\Omega)$ et P la probabilité uniforme. Soit $A = \{1,2\}$, $B = \{1,3\}$ et $C = \{2,3\}$. Alors A, B et C sont deux-à-deux indépendants, mais pas (mutuellement) indépendants.

Passons maintenant au conditionnement. La même approche par les fréquences que ci-dessus conduit à poser:

Définition 13. Soit A et B deux événements, avec $P(A) > 0$. La **probabilité conditionnelle de B si A** est le nombre

$$P(B/A) \;=\; \frac{P(A \cap B)}{P(A)}. \tag{21}$$

Proposition 14: *Supposons que $P(A) > 0$.*

a) A et B sont indépendants si et seulement si $P(B/A) = P(B)$.

b) L'application $B \mapsto P(B/A)$ de $\mathcal{A}$ dans $[0,1]$ est une nouvelle probabilité sur $\mathcal{A}$, appelée la probabilité conditionnelle si A.

Preuve. (a) est évident si on compare (19) et (21). On a $0 \leq P(B/A) \leq 1$. Enfin les propriété (P1) et (P3) pour $P(\bullet/A)$ proviennent des mêmes propriétés pour P et des remarques suivantes: on a $\Omega \cap B = B$, et $(\cup_n B_n) \cap A = \cup_n (B_n \cap A)$, et si B et C sont disjoints il en est de même de $B \cap A$ et $C \cap A$. $\square$

Proposition 15 ("Théorème des probabilités composées"): *Si $A_1, \ldots, A_n$ sont des événements tels que $P(A_1 \cap \ldots \cap A_{n-1}) > 0$, on a*

$$P(A_1 \cap \ldots \cap A_n) \;=\; P(A_1)P(A_2/A_1)P(A_3/A_1 \cap A_2) \ldots P(A_n/A_1 \cap \ldots \cap A_{n-1}). \tag{22}$$

Preuve. On fait une démonstration par récurrence. Si $n = 2$, les formules (22) et (21) sont identiques. Supposons (22) vraie pour $n-1$, et soit $B = A_1 \cap \ldots \cap A_{n-1}$. D'après (21) on a $P(B \cap A_n) = P(B)P(A_n/B)$; en remplaçant $P(B)$ par sa valeur donnée par (22) avec $n-1$, on obtient (22) pour n. $\square$

Proposition 16: *Soit $(B_i)_{i \in I}$ une partition finie ou dénombrable (i.e. l'ensemble d'indices I est fini ou dénombrable) de Ω, constituée d'événements vérifiant $P(B_i) > 0$. Pour tout $A \in \mathcal{A}$ on a alors*

$$P(A) \;=\; \sum_{i \in I} P(A/B_i)P(B_i). \tag{23}$$

Preuve. On a $A = \cup_{i \in I}(A \cap B_i)$, les $A \cap B_i$ sont deux-à-deux disjoints, et $P(A \cap B_i) = P(A/B_i)P(B_i)$. Il suffit alors d'appliquer (P3). $\square$

Proposition 17 ("Théorème de Bayes", ou "de probabilité des causes"): *Sous les mêmes hypothèses que dans la proposition 16, et si $P(A) > 0$, on a*

$$P(B_i/A) \;=\; \frac{P(A/B_i)P(B_i)}{\sum_{j \in I} P(A/B_j)P(B_j)}. \tag{24}$$

Preuve. Le dénominateur de (24) vaut $P(A)$ d'après (23), tandis que (21) implique

$$P(B_i/A) \;=\; \frac{P(A \cap B_i)}{P(A)} \;=\; \frac{P(A/B_i)P(B_i)}{P(A)}. \qquad \square$$

CHAPITRE 2

Probabilités sur les ensembles finis ou dénombrables

1 Quelques résultats utiles sur les séries

Dans ce paragraphe, nous faisons un résumé, essentiellement sans démonstrations, des résultats sur les séries qui sont d'usage constant dans l'étude des probabilités sur un espace dénombrable.

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

$$+\infty+\infty=+\infty, \quad -\infty-\infty=-\infty, \quad a+\infty=+\infty, \quad a-\infty=-\infty \text{ si } a\in\mathbb{R}, \tag{1}$$

$$0\times\infty=0, \quad a\in]0,\infty] \Rightarrow a\times\infty=+\infty, \quad a\in[-\infty,0[\Rightarrow a\times\infty=-\infty. \tag{2}$$

Soit $(u_n)_{n\geq 1}$ une suite numérique, et $S_n=u_1+...+u_n$ la "somme partielle" à l'ordre n.

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

S2 Si la série $\sum_n u_n$ converge, la *suite* $(u_n)_{n\geq 1}$ tend vers 0. La réciproque est **fausse**: on peut avoir $u_n\to 0$ sans que la série $\sum_n u_n$ converge.

S3 La série $\sum_n u_n$ est dite *absolument convergente* si la série $\sum_n |u_n|$ converge.

S4 Si on a $u_n\geq 0$ pour tout n, la suite S_n est croissante, donc elle tend toujours vers une limite $S\in\overline{\mathbb{R}}_+$. On écrit encore $S=\sum_n u_n$, bien que la série converge au sens de (S1) si et seulement si $S<\infty$. Avec les conventions (1) ceci s'applique même si les u_n sont à valeurs dans $\overline{\mathbb{R}}_+$.

En général l'ordre dans lequel on considère les termes d'une série est important. Il existe en effet de nombreux exemples de suites $(u_n)_{n\geq 1}$ et de bijections v de $\mathbb{N}^*$ dans lui-même pour lesquels $\sum_n u_n$ converge et $\sum_n u_{v(n)}$ diverge, ou converge vers une somme différente. Cela étant, il existe deux cas importants où l'ordre des termes n'a pas d'importance:

S5 Lorsque les u_n sont des réels de signe quelconque et lorsque la série est absolument convergente, on peut modifier de manière arbitraire l'ordre des termes sans changer la propriété d'être absolument convergente, ni la somme de la série.

S6 Si $u_n \in \overline{\mathbb{R}}_+$ pour tout n, la somme $\sum_n u_n$ (finie ou infinie: cf. (S4) ci-dessus) ne change pas si on change l'ordre de sommation. Rappelons rapidement la démonstration de cette propriété, qui est fondamentale pour les probabilités: soit v une bijection de $\mathbb{N}^*$ dans lui-même, $S_n = u_1 + \ldots + u_n$ et $S'_n = u_{v(1)} + \ldots + u_{v(n)}$; les suites (S_n) et (S'_n) sont croissantes, et on note S et S' leur limites respectives (dans $\overline{\mathbb{R}}_+$). Pour tout n il existe un entier $m(n)$ tel que $v(i) \leq m(n)$ dès que $i \leq n$; comme $u_i \geq 0$, on a donc clairement $S'_n \leq S_{m(n)} \leq S$, donc en passant à la limite on obtient $S' \leq S$. On montre de même que $S \leq S'$, donc $S = S'$.

S7 Si $u_n \in \overline{\mathbb{R}}_+$, on peut "sommer par paquets". Cela signifie la chose suivante: soit $(A_i)_{i \in I}$ une partition de $\mathbb{N}^*$, avec $I = \{1, 2, \ldots, N\}$ pour un entier N, ou $I = \mathbb{N}^*$. Pour chaque $i \in I$ on pose $v_i = \sum_{n \in A_i} u_n$: si A_i est fini, c'est une somme ordinaire; sinon, v_i est elle-même la somme d'une série à termes positifs. On a alors la propriété que $\sum_n u_n = \sum_{i \in I} v_i$ (cette dernière somme est de nouveau la somme d'une série à termes positifs si $I = \mathbb{N}^*$). La démonstration de ce résultat est tout-à-fait analogue à celle de (S6) ci-dessus.

S8 Si la série $\sum_n u_n$ est absolument convergente, on a la même propriété (S7) ci-dessus.

2 Construction des probabilités sur un ensemble fini ou dénombrable

Dans ce chapitre, on suppose toujours que l'espace d'états Ω est fini ou dénombrable. La tribu des événements est toujours prise égale à l'ensemble $\mathcal{P}(\Omega)$ de toutes les parties de Ω.

Proposition 1: *a) Une probabilité P sur Ω fini ou dénombrable est entièrement caractérisée par ses valeurs sur les singletons, soient $p_\omega = P(\{\omega\})$.*

b) Etant donnée une famille $(p_\omega)_{\omega \in \Omega}$ de réels, il lui correspond une probabilité P (nécessairement unique) telle que $P(\{\omega\}) = p_\omega$ pour tout $\omega \in \Omega$ si et seulement si

$$p_\omega \geq 0, \qquad \sum_{\omega \in \Omega} p_\omega = 1. \tag{3}$$

et dans ce cas, on a pour tout $A \in \Omega$:

$$P(A) = \sum_{\omega \in A} p_\omega \qquad \text{(avec la convention} \quad \sum_{\emptyset} = 0 \text{).} \tag{4}$$

Preuve. Lorsque Ω est fini, ce résultat n'est autre que la proposition 1-2. Lorsque Ω est dénombrable, la démonstration est analogue, si ce n'est que pour prouver que P défini par (4) vérifie (P3), il faut utiliser la propriété de sommation par paquets (S7). $\square$

Exemples:

1) La **loi de Poisson** de paramètre $\theta > 0$ est la probabilité sur $\Omega = \mathbb{N}$ caractérisée par
$$p_n \;=\; e^{-\theta}\,\frac{\theta^n}{n!} \tag{5}$$
(elle vérifie bien (3)). On la note: Poisson(θ). Par extension, si $\theta = 0$, et avec la convention $0^0 = 1$, la formule (5) définit encore une probabilité avec $p_0 = 1$ et $p_n = 0$ pour tout $n \geq 1$.

2) **La loi de Poisson comme limite de lois binomiales**. Soit
$$p_j(a_n, n) \;=\; \begin{cases} C_n^j\;(a_n)^j(1-a_n)^{n-j} & \text{si } \; j = 0, 1, \ldots, n \\ 0 & \text{si } \; j \geq n+1 \end{cases}$$
avec $a_n \in [0, 1]$. Ainsi, pour n fixé, la famille $(p_j(a_n, n))_{j\in\mathbb{N}}$ est l'extension naturelle sur $\Omega = \mathbb{N}$ de la loi binomiale de paramètre a_n et de taille n. Supposons alors que $na_n \to \theta \in \mathbb{R}_+$ lorsque $n \to \infty$. En développant les combinaisons C_n^j, il est facile de vérifier que
$$p_j(a_n, n) \;\to\; p_j \;=\; e^{-\theta}\,\frac{\theta^j}{j!} \qquad \forall j \in \mathbb{N}. \tag{6}$$
Ce résultat est très utile pour les calculs numériques: si n est grand, $p_j(a_n, n)$ est difficile à calculer. Mais si a_n est "petit" on peut remplacer la loi binomiale par la loi de Poisson, ce qui conduit à des calculs plus simples. On verra plus loin une autre approximation de $B(a_n, n)$ lorsque n est grand et a_n n'est pas petit.

3) La **loi géométrique** de paramètre $a \in]0, 1[$ est la probabilité sur $\Omega = \mathbb{N}$ caractérisée par
$$p_n \;=\; (1-a)a^n. \tag{7}$$
Là encore on a (3). On remarquera que si $a \geq 1$, les p_n ci-dessus ne vérifient pas (3).

3 Espérance des variables aléatoires

On suppose donnée une probabilité P sur Ω, caractérisée par les $p_\omega = P(\{\omega\})$. Soit X une variable aléatoire , i.e. une application de Ω dans un espace E. L'image E' de Ω par X (i.e. l'ensemble des $X(\omega)$ lorsque ω parcourt Ω) est nécessairement finie ou dénombrable. La loi de X (cf. (1-7)) est alors la probabilité sur E' caractérisée par les nombres
$$p_j^X \;=\; P(X = j) \;=\; \sum_{\omega : X(\omega) = j} p_\omega, \qquad \forall j \in E'. \tag{8}$$

Définition 2: Soit X une variable aléatoire réelle sur l'espace fini ou dénombrable Ω (i.e. une application de Ω dans $\mathbb{R}$). Son **espérance mathématique**, ou simplement **espérance**, ou parfois **moyenne**, est le nombre
$$E(X) \;=\; \sum_{\omega\in\Omega} p_\omega\; X(\omega), \tag{9}$$

pourvu que la somme $\sum_\omega p_\omega |X(\omega)|$ soit finie: rappelons que, en vertu de (S6), cette somme ne dépend pas de la manière dont les ω sont ordonnés.

Cette définition est motivée ainsi: répétons n fois l'expérience aléatoire, et notons $X_1, \dots, X_n$ les valeurs successives prises par X. Soit aussi $M_n = \frac{1}{n}(X_1 + \dots + X_n)$ leur moyenne arithmétique. On a clairement

$$M_n \;=\; \sum_{\omega \in \Omega} f_n(\{\omega\})\; X(\omega),$$

où $f_n(\{\omega\})$ est la fréquence de réalisation du singleton $\{\omega\}$ au cours des n expériences. Si alors la propriété (1-1) est vraie, c'est-à-dire si $f_n(\{\omega\}) \to p_\omega$, et si dans l'expression ci-dessus on peut intervertir la somme et la limite (ce qui est certainement vrai si Ω est fini), alors la suite M_n tend vers $E(X)$: l'espérance mathématique, ou moyenne, est la limite des moyennes arithmétiques lorsque le nombre d'expériences tend vers l'infini. Nous justifierons cette assertion plus loin, dans l'un des théorèmes les plus importants des probabilités, appelé la loi des grands nombres.

On note $\mathcal{L}^1$ l'ensemble de toutes les variables aléatoires réelles X qui admettent une espérance, c'est-à-dire telles que $\sum_\omega p_\omega |X(\omega)| < \infty$. Cet ensemble dépend évidemment de Ω et de la probabilité P. Les propriétés suivantes sont immédiates:

$$\mathcal{L}^1 \text{ est un espace vectoriel, et l'espérance est linéaire sur } \mathcal{L}^1. \tag{10}$$

$$X \in \mathcal{L}^1 \;\Leftrightarrow\; |X| \in \mathcal{L}^1, \text{ et dans ce cas } \; |E(X)| \leq E(|X|). \tag{11}$$

$$\text{L'espérance est positive: si } X \geq 0 \text{ et } X \in \mathcal{L}^1, \text{ alors } E(X) \geq 0. \tag{12}$$

$$\mathcal{L}^1 \text{ contient toutes les variables aléatoires bornées} \tag{13}$$

(rappelons que X est dite bornée s'il existe un réel b tel que $|X(\omega)| \leq b$ pour tout ω).

$$\text{Si } \; X(\omega) = a \; \text{ pour tout } \omega, \text{ alors } \; E(X) = a. \tag{14}$$

$$\text{Si } \Omega \text{ est fini, } \mathcal{L}^1 \text{ contient toutes les variables aléatoires réelles.} \tag{15}$$

$$\text{Si } A \subset \Omega \text{ et } X = 1_A \text{ (indicatrice de } A\text{), on a } \; E(X) = P(A). \tag{16}$$

On va maintenant introduire un second ensemble de variables aléatoires , l'ensemble $\mathcal{L}^2$ des X réelles telles que le carré X^2 soit dans $\mathcal{L}^1$.

Proposition 3: *$\mathcal{L}^2$ est un sous-espace vectoriel de $\mathcal{L}^1$, et si $X \in \mathcal{L}^2$ on a*

$$|E(X)| \;\leq\; E(|X|) \;\leq\; \sqrt{E(X^2)}. \tag{17}$$

Preuve. Soit X et Y deux variables aléatoires réelles et $a \in I\!R$. Comme $(aX+Y)^2 \leq 2a^2X^2 + 2Y^2$, si X et Y sont dans $\mathcal{L}^2$, on déduit de (10) que $aX + Y \in \mathcal{L}^2$: ainsi $\mathcal{L}^2$ est un espace vectoriel. L'inclusion $\mathcal{L}^2 \subset \mathcal{L}^1$ découle de $|X| \leq 1 + X^2$ et de (10), (11) et (13).

La première inégalité (17) a déjà été vue en (11). Pour la seconde, on peut se contenter du cas où X est positive. Soit alors $a = E(X)$ et $Y = X - a$. D'après (10) il vient

$$E(Y^2) \;=\; E(X^2) - 2aE(X) + a^2 \;=\; E(X^2) - a^2,$$

et $E(Y^2) \geq 0$ par (12). Donc $a^2 \leq E(X^2)$, ce qui est le résultat cherché. $\square$

Définition 4: Si $X \in \mathcal{L}^2$, sa **variance** est l'espérance de la variable aléatoire $[X - E(X)]^2$, et on la note σ^2, ou σ_X^2, ou $\mathrm{var}(X)$. En vertu de (12) elle est positive, et sa racine carrée positive σ s'appelle **l'écart-type** de X.

En développant le carré $[X - E(X)]^2$ comme dans la preuve ci-dessus, on voit aussi que

$$\sigma^2 \;=\; E(X^2) - E(X)^2. \tag{18}$$

Proposition 5 (inégalité de Bienaymé-Tchebicheff): *Soit $X \in \mathcal{L}^2$ de variance σ^2 et $a > 0$. On a alors*

$$P(|X| \geq a) \;\leq\; \frac{E(X^2)}{a^2}, \qquad P(|X - E(X)| \geq a) \;\leq\; \frac{\sigma^2}{a^2}. \tag{19}$$

Preuve. En utilisant (9) et (4), on obtient

$$E(X^2) = \sum_{\omega} p_\omega \; X(\omega)^2 \geq a^2 \sum_{\omega : |X(\omega)| \geq a} p_\omega = a^2 \; P(|X| \geq a),$$

d'où la première inégalité (19). La seconde s'obtient en appliquant la première à la variable aléatoire $X - E(X)$ $\square$

Remarquons aussi que l'espérance $E(X)$ de la variable aléatoire $X \in \mathcal{L}^1$ **ne dépend que de la loi** de X. En effet si E' désigne l'ensemble des valeurs prises par X, on a avec la notation (8):

$$E(X) \;=\; \sum_{\omega \in \Omega} p_\omega X(\omega) \;=\; \sum_{i \in E'} \sum_{\omega : X(\omega) = i} p_\omega i \;=\; \sum_{i \in E'} i p_i^X, \tag{20}$$

où la sommation par paquets est justifiée par (S8) puisque $\sum_\omega \; p_\omega |X(\omega)| < \infty$ par le fait que $X \in \mathcal{L}^1$.

Plus généralement, si X est une variable aléatoire à valeurs dans E, si E' désigne l'ensemble des valeurs prises par X, et si f est une fonction de E' dans $\mathbb{R}$, alors $Y = f(X)$ est une variable aléatoire réelle. On peut aussi considérer f comme une variable aléatoire sur l'espace E' muni de la probabilité P_X (la loi de X). On a alors le résultat fondamental suivant, qui montre la cohérence de la notion d'espérance.

Proposition 6: *Avec les hypothèses précédentes, la variable aléatoire $Y = f(X)$ sur (Ω, P) est dans $\mathcal{L}^1$ si et seulement si la variable aléatoire f sur (E', P_X) est dans $\mathcal{L}^1$;*

dans ce cas, les espérances de ces deux variables aléatoires sont égales, et on a en particulier

$$E(Y) \;=\; E(f(X)) \;=\; \sum_{\omega\in\Omega} f(X(\omega))p_\omega \;=\; \sum_{i\in E'} f(i)p_i^X. \tag{21}$$

Preuve. Comme on peut sommer par paquets par (S7) dans une série à termes positifs, on voit comme pour (20) que les deux expressions de droite de (21) sont égales si on remplace f par $|f|$; elles sont donc finies simultanément, donc d'après la définition de $\mathcal{L}^1$ on a $f(X) \in \mathcal{L}^1(\Omega, P) \;\Leftrightarrow\; f \in \mathcal{L}^1(E', P_X)$.

Si ces propriétés sont réalisées, en utilisant cette fois (S8) on voit de la même manière que les deux expressions de droite de (21) sont aussi égales pour f, ce qui, compte-tenu de (9), achève la démonstration.

Exemples.

1) Soit X une variable aléatoire de loi binomiale $B(p,n)$. On veut calculer son espérance m et sa variance σ^2. A cet effet, on pose

$$g(x) = (1-p+px)^n = \sum_{i=0}^{n} C_n^i\; p^i x^i (1-p)^{n-i}.$$

g est un polynôme, et en le dérivant deux fois en $x=0$ on trouve

$$pn = g'(0) = \sum_{i=0}^{n} i\; C_n^i\; p^i x^i (1-p)^{n-i} = E(X),$$

$$p^2 n(n-1) = g''(0) = \sum_{i=0}^{n} i(i-1)\; C_n^i\; p^i x^i (1-p)^{n-i} = E[X(X-1)]$$

d'après (21). Par suite on a

$$\left.\begin{aligned} m \;&=\; np, \\ \sigma^2 \;&=\; E[X(X-1)] + E(X) - E(X)^2 \;=\; np(1-p). \end{aligned}\right\} \tag{22}$$

2) Soit X une variable aléatoire suivant la loi de Poisson de paramètre θ. On a

$$E(X) \;=\; \sum_{n=0}^{\infty} n\; e^{-\theta}\, \frac{\theta^n}{n!} \;=\; \theta. \tag{23}$$

3) Soit P la loi géométrique de paramètre a sur $\Omega = \mathbb{N}$, et X la variable aléatoire définie par $X(n) = b^n$. Alors, si $|ab| < 1$ on a

$$E(X) \;=\; \sum_{n=0}^{\infty} b^n (1-a) a^n \;=\; \frac{1-a}{1-ab},$$

et si $|ab| \geq 1$ la variable aléatoire X n'a pas d'espérance (i.e. n'est pas dans $\mathcal{L}^1$).

4 Fonction génératrice d'une variable aléatoire à valeurs entières

Dans ce paragraphe on considère une variable aléatoire X à valeurs dans $I\!N$, dont la loi est caractérisée par les nombres $p_n = p_n^X = P(X = n)$.

Définition 7. La **fonction génératrice** de X est la fonction définie sur l'intervalle $[0, 1]$ par la formule suivante (rappelons (21)):

$$g(s) \;=\; E(s^X) \;=\; \sum_{n=0}^{\infty} p_n s^n. \tag{24}$$

Comme on le voit ci-dessus, la fonction génératrice ne dépend que de la loi de X: on parle donc aussi de la fonction génératrice d'une loi, ou d'une probabilité, sur $I\!N$.

Proposition 8: *La fonction génératrice est continue sur $[0, 1]$ et indéfiniment dérivable sur $[0, 1[$; elle caractérise la loi de X.*

Preuve. La fonction g est la somme d'une série entière qui converge absolument au point 1 à cause de (3) ou de (12). Les propriétés de continuité et de dérivabilité de l'énoncé sont alors bien connues. Comme la dérivée $n^{\text{ème}}$ en 0 est $g^{(n)}(0) = p_n n!$, la fonction g caractérise les p_n, donc la loi de X. $\square$

Proposition 9: *Soit X une variable aléatoire à valeurs entières, de fonction génératrice g. Pour que $X \in \mathcal{L}^1$ il faut et il suffit que g soit dérivable à gauche en $s = 1$, et dans ce cas $E(X) = g'(1)$.*

Preuve. Rappelons d'abord un résultat facile sur les séries: si les $s \mapsto u_n(s)$ sont des fonctions croissantes sur $[0, 1[$, positives, on peut échanger limite en s au point 1 et somme en n:

$$\lim_{s\uparrow 1} \sum_n u_n(s) \;=\; \sum_n \lim_{s\uparrow 1} u_n(s). \tag{25}$$

Si $s < 1$ on a

$$\frac{g(s) - g(1)}{s - 1} \;=\; \sum_n p_n \frac{s^n - 1}{s - 1} \;=\; \sum_n p_n (1 + s + \ldots + s^{n-1}),$$

et les fonctions $u_n(s) = p_n(1+s+\ldots+s^{n-1})$ sont croissantes et positives, avec $\lim_{s\uparrow 1} u_n(s) = np_n$. Le résultat découle alors de (25). $\square$

Plus généralement, la même démonstration prouve que la variable aléatoire $X(X - 1)\ldots(X - p)$, est dans $\mathcal{L}^1$ si et seulement si g est $p + 1$ fois dérivable à gauche en $s = 1$, et on a alors

$$E[X(X - 1)\ldots(X - p)] \;=\; g^{(p+1)}(1). \tag{26}$$

Pour se rappeler cette formule, on peut dériver formellement la série (24) terme à terme, $p + 1$ fois, au point $s = 1$, ce qui donne

$$g^{(p+1)}(1) \;=\; \sum_n p_n n(n - 1)\ldots(n - p),$$

et le membre de droite ci-dessus égale le membre de gauche de (26) lorsque ce dernier existe, d'après (21). On peut aussi bien dériver $p+1$ fois les deux membres de l'égalité $g(x) = E(s^X)$, en échangeant les dérivées et le signe espérance (là encore, c'est une manipulation "formelle").

Exemples.

1) Loi de Poisson de paramètre θ: on a

$$g(s) \;=\; e^{\theta(s-1)}. \tag{27}$$

En dérivant, on retrouve (23), i.e. $E(X) = \theta$. En dérivant deux fois, (26) donne $E[X(X-1)] = \theta^2$, donc la variance de X est

$$\sigma^2 \;=\; E[X(X-1)] + E(X) - E(X)^2 \;=\; \theta. \tag{28}$$

2) Loi binomiale $B(p,n)$: on a

$$g(s) \;=\; (1-p+ps)^n \tag{29}$$

(on retrouve la fonction utilisée pour calculer la moyenne et la variance d'une variable aléatoire binomiale dans l'exemple du paragraphe 2).

3) Loi géométrique de paramètre a: on a $g(s) = \frac{1-a}{1-as}$.

5 Variables aléatoires indépendantes

Dans ce paragraphe on considère deux variables aléatoires X et Y définies sur le même espace Ω fini ou dénombrable, muni de la probabilité P. On suppose X et Y à valeurs respectivement dans E et F, et on a vu plus haut qu'on peut toujours supposer que E et F sont eux-mêmes finis ou dénombrables. On pose $p_i^X = P(X=i)$ pour $i \in E$, et $p_i^Y = P(Y=i)$ pour $i \in F$.

On peut aussi considérer le couple $Z = (X,Y)$ comme une variable aléatoire à valeurs dans le produit cartésien $G = E \times F$, et on note sa loi $p_k^Z = P(Z=k)$ pour $k = (i,j) \in G$. On définit enfin la **loi conditionnelle** de Y si $X = i$ par

$$p_j^{Y/X=i} \;=\; P(Y=j/X=i) \qquad \text{si} \;\; p_i^X > 0. \tag{30}$$

Proposition 10: *Il est équivalent de connaitre les $(p_k^Z : k \in G)$ d'une part, les $(p_i^X : i \in E)$ et les $(p_j^{Y/X=i} : j \in F)$ pour les $i \in E$ tels que $p_i^X > 0$ d'autre part, via les formules:*

$$p_i^X \;=\; \sum_{j \in F} p_{(i,j)}^Z, \tag{31}$$

$$p_j^{Y/X=i} \;=\; \frac{p_{(i,j)}^Z}{p_i^X} \qquad si \;\; p_i^X > 0, \tag{32}$$

$$p^Z_{(i,j)} = \begin{cases} p^X_i \; p^{Y/X=i}_j & si \;\; p^X_i > 0 \\ 0 & sinon. \end{cases} \tag{33}$$

Preuve. Il suffit de montrer les trois formules de l'énoncé. D'abord, l'ensemble $\{X = i\}$ est la réunion (finie ou dénombrable) des ensembles deux-à-deux disjoints $\{X = i, Y = j\} = \{Z = (i,j)\}$ pour $j \in F$, donc (31) découle de l'axiome (P3). (32) vient de la formule (1-21). Enfin (33) découle de (32) si $p^X_i > 0$, tandis que si $p^X_i = P(X = i) = 0$ on a *a fortiori* $P(X = i, Y = j) = p^Z_{(i,j)} = 0$. □

Définition 11. Les variables aléatoires X et Y sont dites **indépendantes** si pour toutes parties $A \subset E$, $B \subset F$ on a

$$P(X \in A, Y \in B) \;=\; P(X \in A)P(X \in B). \tag{34}$$

Proposition 12: *Il y a équivalence entre:*

(i) Les variables aléatoires X et Y sont indépendantes.

(ii) On a $p^Z_{(i,j)} = p^X_i p^Y_j$ pour tous $i \in E$, $j \in F$.

(iii) On a $p^{Y/X=i}_j = p^Y_j$ pour tout $j \in F$ et tout $i \in E$ tel que $p^X_i > 0$.

(iii) signifie que la loi conditionnelle de Y sachant $X = i$ est égale à la loi *a priori* de Y, ce qui correspond bien à l'idée intuitive d'indépendance. Bien entendu, comme la définition 11 de l'indépendance est symétrique en X et Y, on peut ci-dessus (comme d'ailleurs dans la proposition 10) échanger les variables aléatoires X et Y.

Preuve. Pour obtenir (i)⇒(ii) il suffit de prendre $A = \{i\}$ et $B = \{j\}$ dans (34). Inversement, supposons (ii). En sommant par paquets dans une série à termes positifs, on obtient pour $A \subset E$ et $B \subset F$:

$$\begin{aligned} P(X \in A, Y \in B) \;&=\; P(Z \in A \times B) \;=\; \sum_{(i,j)\in A\times B} p^Z_{(i,j)} \\ &= \sum_{i\in A}\sum_{j\in B} p^X_i p^Y_j \;=\; \sum_{i\in A} p^X_i \sum_{j\in B} p^Y_j \;=\; P(X \in A)P(Y \in B), \end{aligned}$$

donc on a (i). Enfin, l'équivalence (ii)⇔(iii) provient de (32) et (33) . □

Proposition 13: *Supposons les variables aléatoires X et Y indépendantes, et soit f et g deux fonctions réelles sur E et F respectivement, telles que $f(X) \in \mathcal{L}^1$ et $g(Y) \in \mathcal{L}^1$. Alors le produit $f(X)g(Y)$ est aussi dans $\mathcal{L}^1$, et on a*

$$E[f(X)g(Y)] \;=\; E[f(X)]\; E[g(Y)]. \tag{35}$$

Preuve. Exactement comme dans la démonstration précédente, on peut écrire

$$\sum_{(i,j)\in G} |f(i)g(j)| p^Z_{(i,j)} \;=\; \sum_{i\in E, j\in F} |f(i)g(j)| p^X_i p^Y_j \;=\; \left(\sum_{i\in E} |f(i)| p^X_i\right)\left(\sum_{j\in F} |g(j)| p^Y_j\right),$$

qui est fini par hypothèse: par suite $f(X)g(Y)$ appartient à $\mathcal{L}^1$. En utilisant alors (S8), la même démonstration montre qu'on a les égalités ci-dessus en enlevant les valeurs absolues: cela donne (35). $\square$

Proposition 14: *Supposons que E et F soient contenus dans l'ensemble $\mathbb{Z}$ des entiers relatifs. Soit $U = X + Y$ et $p_i^U = P(U = i)$. Alors*

$$p_i^U \;=\; \sum_{j\in\mathbb{Z}} p^Z_{(j,i-j)} \;=\; \sum_{j\in\mathbb{Z}} p^Z_{(i-j,j)}. \tag{36}$$

En particulier si X et Y sont indépendantes, on a

$$p_i^U \;=\; \sum_{j\in\mathbb{Z}} p_j^X p_{i-j}^Y \;=\; \sum_{j\in\mathbb{Z}} p_{i-j}^X p_j^Y\,. \tag{37}$$

Preuve. (37) suit immédiatement de (36) et de (ii) de la proposition 12. Pour (36), il suffit d'appliquer (P3) et le fait que $\{U = i\}$ est la réunion des ensembles deux-à-deux disjoints $\{X = j, Y = i - j\}$ pour $j \in \mathbb{Z}$, et aussi des $\{X = i - j, Y = j\}$ pour $j \in \mathbb{Z}$. $\square$

Proposition 15: *Supposons les variables aléatoires X et Y indépendantes, à valeurs dans $E = F = \mathbb{N}$, et $U = X + Y$. Notons g_X, g_Y et g_U les fonctions génératrices de X, Y et U. On a alors*

$$g_U \;=\; g_X g_Y. \tag{38}$$

Preuve. Il suffit de remarquer que $g_U(s) = E(s) = E(s^{X+Y})$ et $g_X(s) = E(s^X)$ et $g_Y(s) = E(s^Y)$ pour $s \in [0,1]$, et d'appliquer (35). $\square$

Exemples.

1) Soit X et Y des variables aléatoires indépendantes de lois binomiales respectives $B(p,n)$ et $B(p,m)$ (avec le même paramètre p). D'après (29), $U = X + Y$ vérifie

$$g_U(s) \;=\; (1-p+ps)^n(1-p+ps)^m \;=\; (1-p+ps)^{n+m}.$$

En appliquant encore (29) et la proposition 8, on en déduit que $X + Y$ suit **la loi binomiale** $B(p, n + m)$.

2) Soit X et Y des variables aléatoires indépendantes de lois de Poisson de paramètres respectifs θ et ζ. D'après (27), $U = X + Y$ vérifie

$$g_U(s) \;=\; e^{\theta(s-1)}\, e^{\zeta(s-1)} \;=\; e^{(\theta+\zeta)(s-1)},$$

de sorte que $X + Y$ suit **la loi de Poisson** de paramètre $\theta + \zeta$.

Jusqu'à présent, nous n'avons considéré que des couples de variables aléatoires . Si on a une famille finie $X_1, \ldots, X_n$ de variables aléatoires à valeurs respectivement dans $E_1, \ldots, E_n$, tout ce qui précède s'étend sans difficulté, sauf que les notations deviennent un peu compliquées. La seule chose pouvant peut-être prêter à confusion est la notion d'indépendance; nous le définissons donc ci-dessous:

Définition 16. Les variables aléatoires $X_1, \ldots, X_n$ sont **indépendantes** (ou, "mutuellement indépendantes) si pour toutes parties $A_1 \subset E_1, \ldots, A_n \subset E_n$ on a

$$P(X_1 \in A_1, \ldots, X_n \in A_n) \;=\; \prod_{i=1}^{n} P(X_i \in A_i). \tag{39}$$

Pour que la propriété ci-dessus soit satisfaite il faut et il suffit que, en posant $Z = (X_1, \ldots, X_n)$ (une variable à valeurs dans $G = E_1 \times \ldots \times E_n$, de loi caractérisée par les $(p_k^Z = P(Z = k), k \in G)$), on ait (comme dans la proposition 12):

$$p^Z_{(i_1,\ldots,i_n)} \;=\; \prod_{j=1}^{n} p^{X_j}_{i_j}, \qquad \forall i_j \in E_j. \tag{40}$$

Enfin, si on a une suite **infinie** $(X_n)_{n \in \mathbb{N}^*}$, on pose:

Définition 17. La suite $(X_n)_{n \in \mathbb{N}^*}$ de variables aléatoires est dite **indépendantes** si pour tout n la famille finie $X_1, \ldots, X_n$ est indépendante.

Construction de variables aléatoires de lois données. Jusque là nous sommes partis des variables aléatoires X_i supposées définies sur un certain espace. Souvent il convient d'inverser notre point de vue: on se donne des lois Q_i, chaque Q_i étant définie sur une espace fini ou dénombrable E_i, et on veut construire des variables aléatoires X_i de lois respectives Q_i.

Le problème se pose naturellement, même lorsqu'on a une probabilité Q sur un espace E qui est fini ou dénombrable: existe-t-il une variable de loi Q ? La réponse est positive, et très simple: il suffit de prendre $\Omega = E$, $P = Q$ et l'application "identité" $X(\omega) = \omega$; X est alors une variable aléatoire sur Ω, et sa loi est évidemment Q. Bien entendu, et il s'agit là d'une remarque importante, il existe bien d'autres variables aléatoires qui ont la même loi Q, et qui sont définies sur d'autres espaces et/ou relativement à d'autres probabilités.

Lorsqu'il y a plusieurs lois, sur plusieurs espaces, la situation est plus compliquée. Par exemple, dans le cas où on a deux espaces E et F muni de deux probabilités Q et R, on veut construire deux variables aléatoires X et Y de lois Q et R. Si on construit séparément X et Y comme ci-dessus, il n'y a aucun rapport entre ces deux variables. Si en revanche on veut les construire sur un même espace Ω muni d'une probabilité P, il manque des informations. Pour le voir, considérons le cas où $E = F$ et $Q = R$. Voici alors deux constructions possibles (on note $p_i = Q(\{i\})$ pour $i \in E$):

1) Soit $\Omega = E$, $P = Q$, et $X(\omega) = Y(\omega) = \omega$. On a clairement deux variables X et Y de loi Q, mais bien-sûr $Y = X$ par construction. La loi du couple $Z = (X, Y)$ est alors

$$p^Z_{(i,j)} \;=\; \begin{cases} p_i & \text{si } \; j = i, \\ 0 & \text{sinon.} \end{cases}$$

Les variables aléatoires X et Y ne sont pas indépendantes (intuitivement parlant, c'est même exactement le contraire). Et, avec la notation (30), on a

$$p_j^{Y/X=i} = \begin{cases} 1 & \text{si } j = i, \\ 0 & \text{sinon.} \end{cases}$$

2) Soit $\Omega = E \times E$ et, pour $\omega = (i,j)$, posons $X(\omega) = i$ et $Y(\omega) = j$. Définissons la probabilité P sur Ω par sa valeurs sur les singletons, soit $P(\{(i,j)\}) = p_i p_j$ (cela définit P en vertu de la proposition 1). Dans ce cas, les deux variables aléatoires X et Y sont de loi Q, et elles sont indépendantes (cf. proposition 12).

Il y aurait d'ailleurs bien d'autres manières de construire le couple (X,Y) de sorte que ces deux variables aléatoires aient la loi Q, et que X et Y ne soient pas indépendantes, ni "complètement" dépendantes l'une de l'autre. En fait, ce qui compte, c'est la loi du **couple**, qui donne non seulement la loi de chaque composante X ou Y (par (31), mais aussi la dépendance entre ces deux variables.

On ne peut donc pas donner de résultats généraux. Nous nous contenterons donc du résultat "partiel" suivant, qui est néanmoins très important:

Théorème 18: *Soit $I = \{1,2,\ldots,n\}$ ou $I = \mathbb{N}^*$, et pour chaque $i \in I$ une probabilité Q_i sur un ensemble E_i qui est fini ou dénombrable. Il existe alors un espace d'états Ω muni d'une probabilité P, et des variables aléatoires X_i sur cet espace, telles que chaque X_i soit à valeurs dans E_i, de loi Q_i, et que la famille $(X_i)_{i\in I}$ soit indépendante.*

Preuve. a) Supposons d'abord que $I = \{1,\ldots,n\}$. On pose $\Omega = E_1 \times \ldots \times E_n$ (qui est fini ou dénombrable) et, si $\omega = (k_1,\ldots,k_n) \in \Omega$, $X_i(\omega) = k_i$. Soit aussi P la probabilité (bien définie en vertu de la proposition 1) sur Ω, caractérisée par sa valeur $p_\omega = P(\{\omega\})$ sur chaque singleton $\omega = (k_1,\ldots,k_n)$:

$$p_\omega = \prod_{i=1}^{n} Q_i(\{k_i\})$$

(on a clairement $\sum_{\omega\in\Omega} p_\omega = 1$). Il vient d'après (P3):

$$P(X_i = k_i) = \sum_{\omega: X_i(\omega)=k_i} p_\omega = \sum_{k_l\in E_l, 1\le l\le n, l\neq i} \prod_{r=1}^{n} Q_r(\{k_r\}) = Q_i(\{k_i\}),$$

de sorte que d'une part la loi de X_i est Q_i et que d'autre part on a (40), donc les variables aléatoires $X_1,\ldots,X_n$ sont indépendantes.

b) Dans le cas où $I = \mathbb{N}^*$, les choses sont plus compliquées, puisque l'ensemble $\Omega = \prod_{i\in\mathbb{N}^*} E_i$ est alors infini non dénombrable: on sort donc du cadre de ce chapitre. Le résultat est néanmoins vrai, mais difficile à démontrer, et nous l'admettrons. $\square$

6 Convergence en loi

On a vu aux paragraphes 2 et 1-3 des "convergences de lois". Nous allons ci-dessous formaliser cette notion. On part d'une suite infinie $(X_n)_{n\in\mathbb{N}^*}$ de variables aléatoires , et d'une variable aléatoire X, toutes à valeurs dans le même espace fini ou dénombrable E. En revanche, elles peuvent être définies sur des espaces d'états différents. On caractérise leurs lois par les quantités $p_i^n = P(X_n = i)$ et $p_i = P(X = i)$, pour $i \in E$.

Définition 19. On dit que la suite (X_n) *converge en loi vers* X si pour tout $i \in E$ on a $p_i^n \to p_i$ quand $n \to \infty$. On écrit $X_n \xrightarrow{\mathcal{L}} X$.

On remarquera que cette notion ne fait intervenir que les **lois** des variables X_n et X, de sorte que la terminologie "correcte" devrait être: "les lois des X_n convergent vers celle de X". La convergence en loi n'implique aucune "proximité" des variables aléatoires elles-mêmes.

Proposition 20: *On a $X_n \xrightarrow{\mathcal{L}} X$ si et seulement si $E[f(X_n)] \to E[f(X)]$ pour toute fonction bornée f sur E.*

Preuve. Pour la condition suffisante il suffit de considérer les fonctions indicatrices $f_i = 1_{\{i\}}$: rappelons que $f_i(j) = 1$ si $j = i$ et $f_i(j) = 0$ sinon, de sorte que $E[f_i(X_n)] = p_i^n$ et $E[f_i(X)] = p_i$.

Montrons maintenant la condition nécessaire. Soit f bornée et $a = \sup|f|$. Soit aussi $\varepsilon > 0$. Etant donné (3) il existe une partie finie A de E telle que $\sum_{i\in A} p_i \geq 1 - \varepsilon$. Donc l'hypothèse $p_i^n \to p_i$ pour tout i implique que pour tout n assez grand on a $\sum_{i\in A} p_i^n \geq 1 - 2\varepsilon$, donc

$$|E[f(X)] - \sum_{i\in A} p_i f(i)| \;=\; |\sum_{i\notin A} p_i f(i)| \;\leq\; a\varepsilon,$$

et de même

$$|E[f(X_n)] - \sum_{i\in A} p_i^n f(i)| \;\leq\; 2a\varepsilon.$$

Enfin $\sum_{i\in A} p_i^n f(i) \to \sum_{i\in A} p_i f(i)$ car A est fini. Comme $\varepsilon > 0$ est arbitraire, on en déduit facilement que $E[f(X_n)] \to E[f(X)]$. $\square$

Proposition 21: *Soit X_n, X des variables aléatoires à valeurs dans $\mathbb{N}$, de fonctions génératrices respectives g_n et g. On a $X_n \xrightarrow{\mathcal{L}} X$ si et seulement si $g_n(s) \to g(s)$ pour tout $s \in [0,1]$.*

Preuve. Si $s \in [0,1]$, on a $g_n(s) = E[f(X_n)]$ et $g(s) = E[f(X)]$) avec, pour f, la fonction bornée $f(n) = s^n$. La condition nécessaire découle alors de la proposition précédente.

Pour la réciproque, commençons par un résultat auxiliaire. Soit q_i^n, q_i des réels dans $[0,1]$ et, pour $s \in [0,1[$:

$$\left.\begin{array}{ll} h_n(s) = \sum_{i=1}^{\infty} q_i^n s^i, & h'_n(s) = h_n(s) + q_0^n, \\ h(s) = \sum_{i=1}^{\infty} q_i s^i, & h'(s) = h(s) + q_0. \end{array}\right\} \tag{41}$$

Supposons que $h'_n(s) \to h'(s)$ pour tout $s \in]0,1[$: alors $q_0^n \to q_0$. En effet, on a $0 \leq h_n(s) \leq \sum_{i=1}^{\infty} s^i = \frac{s}{1-s} \leq s$. Par ailleurs pour tout s arbitraire dans $]0,1[$) on a $|h'_n(s) - h'(s)| \leq s$

pour tout n plus grand qu'un certain entier n_s (dépendant de s). Si $n \geq n_s$ on a donc $|q_0^n - q_0| \leq 3s$, et comme s est arbitrairement proche de 0 on en déduit que $q_0^n \to q_0$.

Revenons à notre problème. On suppose que $g_n(s) \to g(s)$ pour tout $s \in [0,1]$, et montrons par récurrence sur i que $p_i^n \to p_i$. Pour $i = 0$ il suffit d'écrire que $p_0^n = g_n(0) \to g(0) = p_0$. Si maintenant on sait que $p_i^n \to p_i$ pour tout $i \leq j$, en posant

$$h'_n(s) = \frac{1}{s^{j+1}} \left(g_n(s) - \sum_{i=0}^{j} p_i^n s^i \right), \qquad h'(s) = \frac{1}{s^{j+1}} \left(g(s) - \sum_{i=0}^{j} p_i s^i \right),$$

on a d'une part $h'_n(s) \to h'(s)$ pour tout $s \in]0,1[$, et d'autre part (41) avec $q_i^n = p_{j+i+1}^n$ et $q_i = p_{j+i+1}$. On déduit alors de notre résultat auxiliaire que $p_{i+1}^n \to p_{i+1}$. $\square$

Exemple: Si les X_n suivent des lois $B(a_n, n)$, avec $na_n \to \theta$ lorsque $n \to \infty$, on a $g_n(s) = (1 - a_n + sa_n)^n$, qui converge vers $g(s) = e^{\theta(s-1)}$: on retrouve le résultat déjà donné dans le paragraphe 2.

CHAPITRE 3

Les variables aléatoires réelles

1 Probabilités sur $I\!R$

Nous avons vu au chapitre 1 la définition générale d'une probabilité P sur un espace quelconque Ω muni d'une tribu $\mathcal{A}$. Un problème fondamental est évidemment de construire, et de caractériser, ces probabilités.

Le paragraphe 2-1 était consacré à la résolution - facile - de ce problème lorsque Ω était fini ou dénombrable. Le cas général - beaucoup plus difficile - fait l'objet de la "théorie de la mesure", ce qui dépasse largement les limites assignées à ce cours.

Nous allons nous contenter de résoudre, sans démonstrations complètes, le cas où $\Omega = I\!R$ (puis le cas $\Omega = I\!R^d$ dans le paragraphe 6), et où la tribu $\mathcal{A}$ est la tribu borélienne $\mathcal{A} = \mathcal{R}$ engendrée par les ouverts, ou par les fermés, ou par les intervalles de la forme $]-\infty, a]$ avec $a \in \mathbb{Q}$ (cf. proposition 1-6). En règle générale, nous omettrons de mentionner la tribu $\mathcal{R}$.

Définition 1. La **fonction de répartition** de la probabilité μ sur $I\!R$ est la fonction suivante:

$$F(x) \;=\; \mu(]-\infty, x]), \qquad x \in I\!R. \tag{1}$$

Proposition 2: *La fonction de répartition F caractérise la probabilité μ sur $I\!R$, et elle vérifie les trois conditions suivantes:*

$$\left.\begin{array}{l} \bullet \text{ elle est croissante} \\ \bullet \text{ elle est continue à droite} \\ \bullet\ \lim_{x\downarrow-\infty} F(x) = 0, \qquad \lim_{x\uparrow+\infty} F(x) = 1. \end{array}\right\} \tag{2}$$

Preuve. Soit $\mathcal{B}$ l'ensemble de toutes les réunions finies d'intervalles deux-à-deux disjoints, de la forme $]x,y]$ avec $-\infty \le x < y < \infty$, ou de la forme $]x,\infty[$ avec $-\infty \le x < \infty$, auxquelles on adjoint l'enemble vide. Il est facile de vérifier que $\mathcal{B}$ est une algèbre, qui d'après la proposition 1-6 engendre la tribu borélienne $\mathcal{R}$.

D'après (1) on a $\mu(]x,y]) = F(y) - F(x)$ si $x < y$. Par suite si $B \in \mathcal{B}$ s'écrit $B = \cup_{i=1}^{n}]x_i, y_i]$ avec $y_i < x_{i+1}$, on a

$$\mu(B) \;=\; \sum_{i=1}^{n} (F(y_i) - F(x_i)).$$

Donc la fonction de répartition caractérise la restriction de la probabilité μ à l'algèbre $\mathcal{B}$. D'après un résultat (difficile) que nous admettrons ici, la connaissance de μ sur $\mathcal{B}$ suffit à déterminer entièrement μ sur la tribu engendrée par $\mathcal{B}$, soit $\mathcal{R}$: on a donc prouvé (!) la première partie de l'énoncé.

La première assertion de (2) est évidente. Pour la seconde, on remarque que si la suite x_n décroît vers x, alors $]-\infty, x_n] \downarrow]-\infty, x]$, donc $F(x_n) \downarrow F(x)$ par la proposition 1-8 et cela entraine la continuité à droite de F. La troisième assertion se montre de manière analogue, en remarquant que $]-\infty, x]$ décroît vers $\emptyset$ (resp. croît vers $I\!R$) lorsque $x \downarrow -\infty$ (resp. $x \uparrow +\infty$). $\square$

Comme F est croissante, elle admet une limite à gauche en chaque point, limite qu'on notera $F(x-)$. En utilisant (2) et la proposition 1-8 on voit donc facilement, exactement comme dans la preuve ci-dessus, que

$$\left.\begin{array}{ll} \mu(]x,y]) = F(y) - F(x), & \mu(]x,y[) = F(y-) - F(x) \\ \mu([x,y]) = F(y) - F(x-), & \mu([x,y[) = F(y-) - F(x-). \end{array}\right\} \qquad (3)$$

En particulier $\mu(\{x\}) = F(x) - F(x-)$ est le "saut" de la fonction F au point x. On a donc $\mu(\{x\}) = 0$ pour tout x si et seulement si la fonction F est continue en tout point.

La proposition 2 admet une réciproque, qui est l'un des résultats les plus difficiles de la théorie de la mesure, et que nous admettrons ici:

Théorème 3: *Si F est une fonction réelle sur $I\!R$, vérifiant les conditions (2), c'est la fonction de répartition d'une (unique) probabilité μ sur $I\!R$ muni de sa tribu borélienne $\mathcal{R}$. On ne peut pas, en général, définir μ sur la tribu $\mathcal{P}(I\!R)$ de toutes les parties de $I\!R$.*

La seconde assertion ci-dessus explique pourquoi, d'un point de vue strictement mathématique, il est absolument nécessaire d'introduire les tribus en probabilités, malgré la complexité que cela engendre. Si on ne le faisait pas, ce qui reviendrait à prendre (sans le dire) la tribu $\mathcal{A} = \mathcal{P}(I\!R)$, il n'existerait que très peu de probabilités sur $I\!R$, à savoir, exactement les "probabilités discrètes" décrites dans l'exemple 3 ci-dessous.

Exemples:

1) Les **mesures de Dirac** (ou: "masses de Dirac"). Soit $a \in I\!R$. On appelle mesure de Dirac en a la probabilité μ sur $I\!R$ qui vérifie

$$\mu(A) \;=\; \begin{cases} 1 & \text{si } \; a \in A \\ 0 & \text{sinon.} \end{cases} \qquad (4)$$

sa fonction de répartition est $F(x) = 1_{[a,\infty[}(x)$.

2) Les probabilités portées par $\mathbb{N}$: comme $\mathbb{N}$ est une partie de $\mathbb{R}$, toute probabilité sur $\mathbb{N}$ peut être considérée comme une probabilité sur $\mathbb{R}$, qui "ne charge que $\mathbb{N}$". Plus précisément, Q étant une probabilité sur $\mathbb{N}$, on définit son "extension" μ à $\mathbb{R}$ en posant $\mu(A) = Q(A \cap \mathbb{N})$. Si $q_n = Q(\{n\})$ pour $n \in \mathbb{N}$, la fonction de répartition F de μ est

$$F(x) = \begin{cases} 0 & \text{si } x < 0 \\ q_0 + \ldots + q_n & \text{si } n \leq x < n+1, \; n \in \mathbb{N}. \end{cases} \tag{5}$$

3) Les **probabilités discrètes**. Plus généralement, si E est une partie finie ou dénombrable de $\mathbb{R}$, toute probabilité Q sur E peut être considérée comme une probabilité μ sur $\mathbb{R}$, via la formule $\mu(A) = Q(A \cap E)$. Si pour tout $i \in E$ on pose $q_i = Q(\{i\})$, la fonction de répartition F de μ est alors

$$F(x) = \sum_{i \in E : i \leq x} q_i \tag{6}$$

avec la convention qu'une somme "vide" vaut 0. On retrouve bien (5) dans le cas où $E = \mathbb{N}$. On voit que F est **purement discontinue**, au sens où elle est complètement caractérisée par ses sauts $\Delta F(x) = F(x) - F(x-)$, via la formule

$$F(x) = \sum_{y \leq x} \Delta F(y).$$

Noter aussi que l'ensemble E, quoiqu'au plus dénombrable, peut tout-à-fait être partout dense dans $\mathbb{R}$, par exemple il peut être égal à l'ensemble des rationnels $\mathbb{Q}$: si alors $q_i > 0$ pour tout $i \in \mathbb{Q}$, la fonction F est discontinue en tout rationnel (et continue partout ailleurs...).

Il existe bien d'autres probabilités, non discrètes, sur $\mathbb{R}$. Le paragraphe suivant est consacré à un exemple très important, celui des probabilités avec densité.

2 Les densités de probabilité

Dans la suite de ce chapitre nous aurons continuellement à considérer des intégrales de fonctions réelles sur $\mathbb{R}$. L'intervalle d'intégration sera en général $\mathbb{R}$ tout entier, ou parfois un intervalle de la forme $]-\infty, x]$: il s'agit donc la plupart du temps d'intégrales "généralisées". On pourra considérer que les fonctions qu'on intègre sont "intégrables au sens de Riemann", bien que tout fonctionne encore si on considère des fonctions "intégrables au sens de Lebesgue". Dans ce qui suit nous parlerons donc simplement de **fonctions intégrables**: cela signifie que la fonction f qu'on intègre est intégrable au sens de Riemann ou de Lebesgue (ce qui est plus général), selon les connaissances préalables du lecteur, et aussi que l'intégrale généralisée $\int_{-\infty}^{+\infty} |f(x)| dx$ existe (donc est finie), ce qui implique aussi que l'intégrale généralisée $\int_{-\infty}^{+\infty} f(x) dx$ existe.

Définition 4. Une fontion réelle f sur $\mathbb{R}$ est une **densité de probabilité**, ou simplement une "densité", si elle est positive, intégrable, et vérifie

$$\int_{-\infty}^{+\infty} f(x) dx = 1. \tag{7}$$

Si f est comme ci-dessus, la fonction

$$F(x) \ = \ \int_{-\infty}^{x} f(y)dy \tag{8}$$

vérifie les propriétés (2). C'est donc la fonction de répartition d'une probabilité μ: on dit que f est la **densité de** μ, ou que μ admet la densité f.

Dans ce cas, la fonction F est continue, de sorte que $\mu(\{x\}) = 0$ pour tout x, et elle est même dérivable et de dérivée f en tout point x où f est continue. A l'inverse, si la fonction de répartition F d'une probabilité μ est dérivable, ou seulement continue partout et dérivable par morceaux, alors μ admet une densité.

Il existe bien-sûr des probabilités sur $I\!R$ qui n'ont pas de densité: c'est le cas des probabilités discrètes données en exemple au paragraphe 1. Il existe des cas "mixtes": soit d'une part f une fonction positive intégrable et une partie finie ou dénombrable E de $I\!R$ et des $q_i > 0$ indicés par $i \in E$, tels que

$$\int_{-\infty}^{+\infty} f(x)dx + \sum_{i\in E} q_i \ = \ 1.$$

Alors, la fonction

$$F(x) \ = \ \int_{-\infty}^{x} f(y)dy + \sum_{i\in E:i\leq x} q_i \tag{9}$$

est une fonction de répartition, et la probabilité associée μ n'admet pas de densité, et n'est pas non plus discrète. Enfin, on peut même trouver des probabilités dont la fonction de répartition est continue, mais qui n'admettent pas de densité.

Remarques:

1) La fonction de répartition est, de manière évidente, entièrement déterminée par la probabilité μ. Il n'en est pas de même de la densité, lorsqu'elle existe: si en effet on a (8) et si on pose $g(x) = f(x)$ si $x \notin E$ et $g(x) = f(x) + 1$ si $x \in E$, où E est un ensemble fini ou dénombrable, alors g est encore une densité de μ.

2) Voici une interprétation "intuitive" de la densité f de μ. Si Δx est un "petit" accroissement de la variable x, on a (si du moins f est continue en x):

$$f(x) \ \sim \ \frac{\mu([x, x + \Delta x])}{\Delta x}. \tag{10}$$

Exemples:

1) **La loi uniforme sur** $[a, b]$: on a ici deux réels $a < b$, et c'est la probabilité μ admettant la densité

$$f(x) \ = \ \begin{cases} \frac{1}{b-a} & \text{si } \ a \leq x \leq b \\ 0 & \text{sinon.} \end{cases} \tag{11}$$

En vertu de la remarque 1 ci-dessus, on aurait aussi bien pu choisir $f(a) = 0$ ou $f(b) = 0$. Au vu de l'interprétation (10), le fait que f soit constante sur $[a, b]$ correspond au fait que si on choisit une point selon la probabilité uniforme on a "autant de chances" de tomber au voisinage de chaque point de l'intervalle $[a, b]$, ce qui explique le nom "uniforme". Remarquer aussi que $\mu(\{x\}) = 0$ pour tout x (comme pour toutes les probabilités avec densité): on a donc une probabilité **nulle** de tomber exactement en un point x fixé à l'avance.

2) **La loi exponentielle de paramètre** $\theta > 0$: c'est la loi de densité

$$f(x) \;=\; \begin{cases} 0 & \text{si } \; x < 0 \\ \theta e^{-\theta x} & \text{sinon,} \end{cases} \tag{12}$$

et de fonction de répartition

$$F(x) \;=\; \begin{cases} 0 & \text{si } \; x < 0 \\ 1 - e^{-\theta x} & \text{sinon,} \end{cases} \tag{13}$$

3) **La loi normale centrée réduite** (ou loi de Gauss): c'est la loi de densité

$$f(x) \;=\; \frac{1}{\sqrt{2\pi}}\, e^{-x^2/2}. \tag{14}$$

Pour vérifier que cette fonction est d'intégrale 1, on remarque que $I = \int_{-\infty}^{+\infty} f(x)dx$ vérifie

$$I^2 \;=\; \int\int_{\mathbb{R}^2} f(x)f(y)dxdy \;=\; \int_0^{2\pi} d\theta \int_0^{\infty} \frac{1}{2\pi}\, e^{-\rho^2/2}\rho d\rho$$

(en passant en coordonnées polaires dans l'intégrale double), et un calcul simple montre alors que $I^2 = 1$.

Nous aurons l'occasion de voir par la suite un grand nombre d'autres exemples de probabilités avec densité.

3 Les variables aléatoires réelles

Commençons par remarquer que la définition des variables aléatoires donnée au chapitre 1 est mathématiquement insuffisante, sauf si l'espace d'états Ω est fini ou dénombrable. En effet, si X est une variable aléatoire réelle, on voudrait pouvoir définir sa loi par la formule (1-7), de façon à obtenir une probabilité sur $\mathbb{R}$; mais pour cela il faut que l'ensemble $X^{-1}(B)$ soit un événement (i.e. appartienne à la tribu $\mathcal{A}$) pour tout $B \in \mathcal{R}$. Cela nous conduit à poser:

Définition 5. Soit l'espace d'état Ω muni de la tribu $\mathcal{A}$ des événements. Une application X de Ω dans $\mathbb{R}$ est une **variable aléatoire** si $X^{-1}(B) \in \mathcal{A}$ pour tout $B \in \mathcal{R}$.

On a alors le résultat très utile suivant:

Proposition 6: *Si $X_1, \ldots, X_n$ sont des variables aléatoires réelles et si f est une fonction continue de $\mathbb{R}^n$ dans $\mathbb{R}$, alors $Y = f(X_1, \ldots, X_n)$ est une variable aléatoire.*

(Le même résultat est d'ailleurs vrai avec des fonctions f bien plus générales: il suffit que $f^{-1}(B)$ soit un borélien de $I\!R^n$ pour tout $B \in \mathcal{R}$).

Preuve. 1) Montrons d'abord un résultat auxiliaire, à savoir que si X est une application de Ω dans $I\!R$ telle que $\{\omega : X(\omega) \leq a\} = X^{-1}(]-\infty, a])$ appartienne à la tribu $\mathcal{A}$ pour tout $a \in I\!R$, alors X est une variable aléatoire.

Pour cela, on note $\mathcal{A}'$ la tribu de Ω engendrée par les ensembles $X^{-1}(]-\infty, a])$ pour $a \in I\!R$, de sorte que $\mathcal{A}' \subset \mathcal{A}$ par hypothèse. Soit aussi $\mathcal{R}'$ l'ensemble des $B \in \mathcal{R}$ tels que $X^{-1}(B)$ soient dans $\mathcal{A}'$. Comme l'image réciproque X^{-1} commute avec la réunion, l'intersection et le passage au complémentaire, il est clair que $\mathcal{R}'$ est une tribu contenue dans $\mathcal{R}$, et elle contient les intervalles $]-\infty, a]$ par construction: vu la proposition 1-6, on a donc $\mathcal{R}' = \mathcal{R}$, ce qui veut dire que $X^{-1}(B) \in \mathcal{A}$ pour tout $B \in \mathcal{R}$, et X est une variable aléatoire.

2) D'après ce qui précède, il suffit de montrer que les ensembles $\{Y \leq a\}$, ou de manière équivalente les ensembles $\{Y > a\}$, sont dans $\mathcal{A}$ pour tout $a \in I\!R$. Or f étant continue, $A = \{x \in I\!R^n : f(x) > a\}$ est un ouvert. Il s'écrit donc comme réunion dénombrable $A = \cup_{i \in I\!N} A_i$ d'ensembles A_i qui sont des "rectangles ouverts" de la forme $A_i = \prod_{j=1}^n]x_{i,j}, y_{i,j}[$, et on a

$$\{Y > a\} = \{(X_1, \ldots, X_n) \in A\} = \cup_i \{(X_1, \ldots, X_n) \in A_i\} = \cup_i \cap_{j=1}^n \{x_{i,j} < X_j < y_{i,j}\}.$$

Comme par hypothèse $\{x_{i,j} < X_j < y_{i,j}\} \in \mathcal{A}$, on en déduit le résultat. $\square$

Comme application de cet énoncé, on a les propriétés suivantes, où X, Y et les $(X_n)_{n \in I\!N^\star}$ sont des variables aléatoires réelles:

$$X + Y, \quad XY, \quad \frac{X}{Y} \text{ si } Y \neq 0 \quad \text{sont des variables aléatoires;} \tag{15}$$

$$\sup_{1 \leq n \leq p} X_n, \quad \inf_{1 \leq n \leq p} X_n \quad \text{sont des variables aléatoires;} \tag{16}$$

$$\sup_{n \geq 1} X_n, \quad \inf_{n \geq 1} X_n \quad \text{sont des variables aléatoires;} \tag{17}$$

(pour le voir on remarque que $\{\sup_i X_i \leq a\} = \cap_i \{X_i \leq a\}$, qui est dans $\mathcal{A}$ par hypothèse, et on applique la partie (1) de la preuve précédente; même chose pour l'inf);

$$X_n(\omega) \to Z(\omega), \quad \forall \omega \quad \Rightarrow \quad Z \text{ est une variable aléatoire} \tag{18}$$

(On a en effet $Z = \inf_n Y_n$, où $Y_n = \sup_{i \geq n} X_i$, et on applique deux fois (17));

$$Z = 1_A \;\text{ (indicatrice de } A\text{) est une variable aléatoire } \Leftrightarrow \; A \in \mathcal{A} \tag{19}$$

(il suffit de remarquer que $Z^{-1}(B)$ égale $\emptyset$, A, A^c ou Ω selon que $B \cap \{0,1\}$ égale $\emptyset$, $\{1\}$, $\{0\}$ ou $\{0,1\}$).

Maintenant que la notion de variable aléatoire réelle est bien établie, on peut définir la loi P_X de X comme la probabilité sur $I\!R$ muni de la tribu borélienne $\mathcal{R}$, et définie par (1-7). On note aussi F_X la fonction de répartition de P_X, qu'on appelle aussi **fonction de répartition de** X; de même si P_X admet une densité f_X, on l'appelle aussi la **densité de** X.

4 Espérance des variables aléatoires réelles

Là encore, nous allons nous contenter de résultats partiels, les résultats complets étant un peu difficiles à démontrer.

Commençons par des considérations générales, qui permettent de comprendre qu'on puisse définir l'espérance de toutes les variables aléatoires réelles “suffisamment petites”, par exemple bornées. On appelle **variable aléatoire étagée** toute variable aléatoire X qui ne prend qu'un nombre fini de valeurs, disons $a_1, \dots, a_p$. D'après (2-19) elle admet une espérance mathématique donnée par

$$E(X) \;=\; \sum_{i=1}^{p} a_i P(X = a_i). \tag{20}$$

Dans une seconde étape, on construit l'espérance des variables aléatoires positives. Si X est une telle variable, on considère une suite X_n de variables aléatoires positives étagées croissant vers X, par exemple

$$X_n(\omega) \;=\; \begin{cases} k2^{-n} & \text{si } \; k2^{-n} \leq X(\omega) < (k+1)2^{-n} \qquad \text{et } \; 0 \leq k \leq n2^n - 1 \\ n & \text{sinon.} \end{cases} \tag{21}$$

Comme $X_n \leq X_{n+1}$, on a $E(X_n) \leq E(X_{n+1})$ par (2-12), et on peut poser

$$E(X) \;=\; \lim_{n\to\infty} E(X_n). \tag{22}$$

Cette limite existe toujours, elle est positive, mais elle peut évidemment être infinie. On peut montrer qu'elle ne dépend pas de la suite (X_n) choisie, pourvu que ce soient des variables aléatoires positives étagées croissant vers X.

Soit enfin X une variable aléatoire de signe quelconque. Elle s'écrit $X = X^+ - X^-$, où $X^+ = \sup(X, 0)$ et $X^- = \sup(-X, 0)$ sont les parties “positive” et “négative” de X, de sorte que $|X| = X^+ + X^-$, et évidemment X^+ et X^- sont deux variables aléatoires positives.

Définition 7. On dit que la variable aléatoire X appartient à $\mathcal{L}^1$ si $E(X^+)$ et $E(X^-)$ sont tous les deux finis. Dans ce cas, l'espérance mathématique de X est le nombre

$$E(X) \;=\; E(X^+) - E(X^-) \qquad \left(\text{noté aussi } \int X(\omega) P(d\omega)\ \right). \tag{23}$$

le lecteur vérifiera sans peine que ces définitions coïncident avec celles du paragraphe 2-2 lorsque Ω est fini ou dénombrable. On montre aussi que **les propriétés 2-(10), (11), (12), (13), (16) sont vraies** dans notre cadre général. Une autre propriété importante est la suivante:

$$\text{si } \; P(X = X') = 1 \;\text{ alors }\; X \in \mathcal{L}^1 \Leftrightarrow X' \in \mathcal{L}^1, \;\text{ et dans ce cas }\; E(X) = E(X') \tag{24}$$

(il suffit clairement de montrer cela lorsque X et X' sont positives; on les approche par des variables étagées X_n et X'_n comme en (21), de sorte que $P(X'_n = X_n) = 1$, donc *a fortiori* $P(X'_n = k2^{-n}) = P(X_n = k2^{-n})$ pour tout k, donc $E(X'_n) = E(X_n)$ par (20).)

Outre l'espace $\mathcal{L}^1$, on utilise aussi, comme au chapitre 2, l'espace $\mathcal{L}^2$ des variables aléatoires X telles que le carré X^2 soit dans $\mathcal{L}^1$.

Proposition 8: *$\mathcal{L}^2$ est un sous-espace vectoriel de $\mathcal{L}^1$; si X et Y sont dans $\mathcal{L}^2$, alors on a (2-17), et $XY \in \mathcal{L}^1$, et l'inégalité de Cauchy-Schwarz:*

$$|E(XY)| \;\leq\; \sqrt{E(X^2)E(Y^2)}. \tag{25}$$

Preuve. La première assertion se montre comme la proposition 2-3, et (2-17) aussi (c'est aussi une conséquence de (25)).

Comme $|XY| \leq \frac{1}{2}(X^2+Y^2)$, on a $XY \in \mathcal{L}^1$ dès que $X, Y \in \mathcal{L}^2$. Enfin, pour tout $x \in I\!R$ on a d'après la linéarité et la positivité de l'espérance:

$$x^2E(X^2) + 2xE(XY) + E(Y^2) \;=\; E[(xX+Y)^2] \;\geq\; 0.$$

Mais ceci n'est possible que si ce trinôme en x n'a au plus qu'une seule racine réelle; son discriminant doit donc être négatif ou nul, ce qui donne immédiatement (25). □

Définition 9. Si $X \in \mathcal{L}^2$, sa **variance** est l'espérance de la variable aléatoire $[X-E(X)]^2$, et on la note σ^2, ou σ_X^2, ou $\mathrm{var}(X)$. Elle est positive, et sa racine carrée positive σ s'appelle **l'écart-type** de X.

On a évidemment encore (2-18).

Définition 10. Si X et Y sont dans $\mathcal{L}^2$, leur **covariance** est l'espérance de la variable aléatoire $[X-E(X)][Y-E(Y)]$, et on la note $\mathrm{cov}(X,Y)$. Le **coefficient de corrélation** des variables aléatoires X et Y est le nombre

$$\rho \;=\; \frac{\mathrm{cov}(X,Y)}{\sqrt{\mathrm{var}(X)\mathrm{var}(Y)}}. \tag{26}$$

Noter que, comme pour (2-18), on a (par le même calcul):

$$\mathrm{cov}(X,Y) \;=\; E(XY) - E(X)E(Y), \tag{27}$$

et d'ailleurs (2-18) est un cas particulier de (27), car $\mathrm{var}(X)$=$\mathrm{cov}(X,X)$. La linéarité de l'espérance donne immédiatement pour $a, b, a', b' \in I\!R$:

$$E[(aX+b)(a'Y+b')] \;=\; aa'E(XY) + ab'E(X) + a'bE(Y) + bb',$$

$$E(aX+b)\;E(a'Y+b') \;=\; aa'E(X)E(Y) + ab'E(X) + a'bE(Y) + bb'.$$

Donc au vu de (27) on a

$$\mathrm{cov}(aX+b, a'Y+b') \;=\; aa'\;\mathrm{cov}(X,Y), \tag{28}$$

et en particulier $\mathrm{var}(aX+b) = a^2\mathrm{var}(X)$. On en déduit que les coefficients de corrélation de X et Y et de $aX+b$ et $a'Y+b'$ sont égaux. Enfin, d'après (25), il vient immédiatement

$$-1 \;\leq\; \rho \;\leq\; 1. \tag{29}$$

Proposition 11 (inégalité de Bienaymé-Tchebicheff): *Soit $X \in \mathcal{L}^2$ de variance σ^2 et $a > 0$. On a alors les inégalités (2-19).*

Preuve. On a $X^2 \geq a^2 1_{[a,\infty[}(|X|)$, donc $E(X^2) \geq a^2 E\left(1_{[a,\infty[}(|X|)\right) = a^2 P(|X| \geq a)$, ce qui donne la première formule (2-19). La seconde découle de la première appliquée à $X - E(X)$. $\square$

Le résultat suivant est l'analogue de la proposition 2-6. On y considère une application de Ω dans un ensemble E muni d'une tribu $\mathcal{E}$, et qui est telle que $X^{-1}(B) \in \mathcal{A}$ pour tout $B \in \mathcal{E}$; ainsi, la "loi" de X, définie par (1-7) pour $B \in \mathcal{E}$, est une probabilité P_X sur la tribu $\mathcal{E}$. Soit également h une application de E dans $I\!R$, telle que $h^{-1}(C) \in \mathcal{E}$ pour tout $C \in \mathcal{R}$: en d'autres termes, h est une "variable aléatoire" sur l'espace E muni de la tribu $\mathcal{E}$. Il est alors évident que $Y = h(X)$ est une variable aléatoire sur Ω avec la tribu $\mathcal{A}$.

Proposition 12: *Sous les hypothèses précédentes, h appartient à $\mathcal{L}^1(E, \mathcal{E}, P_X)$ si et seulement si $h(X)$ appartient à $\mathcal{L}^1(\Omega, \mathcal{A}, P)$, et dans ce cas les espérances de h relativement à P_X et de $h(X)$ relativement à P sont égales. On écrit:*

$$E[h(X)] \;=\; \int h(X(\omega))P(d\omega) \;=\; \int h(x)P_X(dx). \tag{30}$$

Preuve. (30) est évident lorsque h ne prend qu'un nombre fini de valeurs, grâce à (20) et au fait que $P(X^{-1}(B)) = P_X(B)$. Si h et positive, on l'approche par des fonctions étagées h_n comme en (21), et on en déduit encore (30), les trois membres étant simultanément finis ou infinis. Le résultat général s'en déduit par différence. $\square$

Pour terminer ce paragraphe, indiquons comment on peut calculer l'espérance d'une variable aléatoire réelle, ou plus généralement d'une fonction d'une variable aléatoire réelle, lorsque cette dernière admet une densité.

Proposition 13: *Soit X une variable aléatoire réelle admettant la densité f, et soit g une fonction de $I\!R$ dans $I\!R$, continue par morceaux (on sait alors, par une extension triviale de la proposition 6, que $g(X)$ est aussi une variable aléatoire). On a alors $g(X) \in \mathcal{L}^1$ si et seulement si*

$$\int_{-\infty}^{+\infty} |g(x)|f(x)dx \;<\; \infty, \tag{31}$$

et dans ce cas on a

$$E[g(X)] \;=\; \int_{-\infty}^{+\infty} g(x)f(x)dx. \tag{32}$$

Nous n'avons pas les éléments permettant de montrer ce résultat, mais l'argument "heuristique" suivant permet de comprendre pourquoi il est vrai: supposons f et g continues. Posons $X_n = i/n$ si $i/n \leq X < (i+1)/n$, pour $i \in \mathbb{Z}$. Ainsi, $X_n(\omega) \to X(\omega)$ pour tout ω, et par continuité de g on a $g(X_n) \to g(X)$. De plus, comme X_n est une variable aléatoire discrète, on a en utilisant (20) et (10):

$$E[g(X_n)] \;=\; \sum_{i \in \mathbb{Z}} g(\frac{i}{n})P(\frac{i}{n} \leq X < \frac{i+1}{n}) \;\sim\; \sum_{i \in \mathbb{Z}} g(\frac{i}{n})f(\frac{i}{n})\frac{1}{n},$$

et le dernier terme ci-dessus tend vers le second membre de (32) lorsque $n \to \infty$, par approximation de Riemann.

Enfin, sous les mêmes hypothèses sur g, et lorsque la fonction de répartition de X n'admet pas de densité mais s'écrit sous la forme (9), on peut montrer qu'on a le "mélange" suivant de (32) et de (2-21):

$$E[g(X)] = \int_{-\infty}^{+\infty} g(x)f(x)dx + \sum_{i \in E} q_i g(i). \tag{33}$$

5 Exemples de probabilités sur $I\!R$

1) La loi uniforme sur $[a,b]$: Cette probabilité admet la densité donnée par (11). Si la variable aléatoire X admet cette loi, son espérance vaut

$$E(X) = \int_a^b x \frac{1}{b-a}\,dx = \frac{a+b}{2}. \tag{34}$$

2) La loi exponentielle de paramètre $\theta > 0$: Elle admet la densité donnée par (12), et si la variable aléatoire X admet cette loi, sa moyenne et sa variance sont données par

$$m = \frac{1}{\theta}, \qquad \sigma^2 = \frac{1}{\theta^2}. \tag{35}$$

En effet, $m = \int_0^\infty x\theta e^{-\theta x}dx$ et $E(X^2) = \int_0^\infty x^2\theta e^{-\theta x}dx$ et $\sigma^2 = E(X^2) - E(X)^2$, donc deux intégrations par parties successives donnent les quantités ci-dessus. La loi exponentielle jouit aussi d'une propriété importante pour les applications:

Proposition 14 (Propriété de "non-vieillissement"): *Soit X une variable aléatoire positive, telle que $P(X > s) > 0$ pour tout $s \in I\!R$. On a $P(X > t+s/X > t) = P(X > s)$ pour tous $s,t > 0$ si et seulement si X suit une loi exponentielle.*

Preuve. Soit $G(t) = P(X > t) = 1 - F(t)$, où F est la fonction de répartition de X. D'après (1-21), la propriété de l'énoncé équivaut à dire que $G(t+s) = G(t)G(s)$ pour tous $s,t > 0$. Comme G est décroissante et continue à droite et tend vers 0 à l'infini, cela revient aussi à dire que c'est une exponentielle négative, de la forme $G(t) = e^{-\theta t}$ pour un $\theta > 0$. Le résultat s'obtient alors en comparant à (13). $\square$

3) Les lois gamma: Rappelons d'abord que la fonction gamma est définie pour $\alpha \in]0,\infty[$ par

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}dx. \tag{36}$$

Une intégration par parties montre que $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$, et on a de manière évidente $\Gamma(1) = 1$: il s'ensuit que $\Gamma(n+1) = n!$ pour tout entier $n \geq 0$, avec la convention que $0! = 1$. Soit alors $\theta > 0$ et $\alpha > 0$. Le changement de variable $x \mapsto \theta x$ montre alors que la fonction positive suivante

$$f(x) = \begin{cases} 0 & \text{si } x \leq 0 \\ \frac{1}{\Gamma(\alpha)}\theta^\alpha x^{\alpha-1}e^{-\theta x} & \text{sinon} \end{cases} \tag{37}$$

est d'intégrale (sur $\mathbb{R}$) égale à 1. C'est donc la densité d'une probabilité qu'on appelle la loi gamma de paramètre d'échelle θ et d'indice α, et qu'on note $\Gamma(\alpha,\theta)$. On remarque que $\Gamma(1,\theta)$ est la loi exponentielle de paramètre θ.

Si X est une variable aléatoire de loi $\Gamma(\alpha,\theta)$, sa moyenne, sa variance σ^2, et l'espérance $E(X^\beta)$ pour tout $\beta > -\alpha$, sont données par

$$m = \frac{\alpha}{\theta}, \qquad \sigma^2 = \frac{\alpha}{\theta^2}, \qquad E(X^\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)} \frac{1}{\theta^\beta} \tag{38}$$

(utiliser le fait que $E(X^\beta) = \int_0^\infty x^\beta f(x)dx$ et (36)). En revanche si $\beta \leq -\alpha$, on a $E(X^\beta) = +\infty$.

4) La loi normale centrée réduite: La densité de cette loi est donnée par (14). La moyenne m et la variance σ^2 d'une variable aléatoire X admettant cette loi sont

$$m = 0, \qquad \sigma^2 = 1 \tag{39}$$

($m = 0$ car c'est l'intégrale d'une fonction impaire; pour calculer σ^2 on peut faire une intégration par parties). Cette loi est notée $\mathcal{N}(0,1)$, et les qualificatifs "centrée" et "réduite" proviennent précisément de (39).

5) La loi normale $\mathcal{N}(m,\sigma^2)$: Si $m \in \mathbb{R}$ et $\sigma^2 > 0$, c'est la loi de densité

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp -\frac{(x-m)^2}{2\sigma^2}. \tag{40}$$

On voit que f est une densité en se ramenant par le changement de variable $x \mapsto \frac{x-m}{\sigma}$ à la fonction (14). Le même changement de variable permet de voir que m et σ^2 sont la moyenne et la variance d'une variable aléatoire de loi $\mathcal{N}(m,\sigma^2)$.

6 Lois de probabilité sur $\mathbb{R}^n$

De même qu'en dimension 1, une probabilité μ sur $\mathbb{R}^n$, muni de la tribu borélienne $\mathcal{R}^n$, est caractérisée par la fonction de répartition multidimensionnelle $F : \mathbb{R}^n \to \mathbb{R}$ définie par

$$F(x_1,\ldots,x_n) = \mu\left(\prod_{i=1}^n]-\infty, x_i]\right). \tag{41}$$

Mais caractériser les fonctions de répartition sur $\mathbb{R}^n$ est assez délicat, de sorte que cette notion est rarement utilisée. Bien plus utile est la notion de densité, lorsqu'elle existe:

Définition 15. Une fonction réelle sur $\mathbb{R}^n$ est une **densité de probabilité**, ou simplement une "densité", si elle est positive sur $\mathbb{R}^n$, intégrable, et vérifie

$$\int_{\mathbb{R}^n} f(x)dx = \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} f(x_1,\ldots,x_n)dx_1 \ldots dx_n = 1. \tag{42}$$

Si f est une fonction comme ci-dessus, il existe une probabilité μ sur $\mathbb{R}^n$ (muni de la tribu borélienne $\mathcal{R}^n$) et une seule, telle que

$$\mu\left(\prod_{i=1}^{n}]-\infty,x_i]\right) \;=\; \int_{-\infty}^{x_1}\cdots\int_{-\infty}^{x_n} f(y_1,\ldots,y_n)dy_1\ldots dy_n. \tag{43}$$

(on écrit aussi $=\int_{\prod_{i=1}^{n}]-\infty,x_i]} f(y)dy$). On dit alors que f est une densité de μ, ou que μ admet la densité f. Attention: comme dans le cas $n=1$, il existe des probabilités sur $I\!R^n$ qui n'admettent pas de densité.

Une variable aléatoire X à valeurs dans $I\!R^n$ est simplement une collection de n variables réelles, qui sont les "composantes" de X: on écrit $X=(X_1,\ldots,X_n)$. La loi P_X de X est la probabilité définie sur les boréliens $B\in\mathcal{R}^n$ par (1-7), et on dit que X admet la densité f si la fonction f est une densité de P_X au sens ci-dessus. Exactement comme pour la proposition 13, on a:

Proposition 16: *Soit X une variable aléatoire à valeurs dans $I\!R^n$, admettant la densité f, et soit g une fonction de $I\!R^n$ dans $I\!R$, continue par morceaux (i.e. continue sauf sur une "bonne" surface de dimension au plus $n-1$). On a alors $g(X)\in\mathcal{L}^1$ si et seulement si*

$$\int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty} |g(x_1,\ldots,x_n)|f(x_1,\ldots,x_n)dx_1\ldots dx_n \;<\; \infty, \tag{44}$$

et dans ce cas on a

$$E[g(X)] \;=\; \int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty} g(x_1,\ldots,x_n)f(x_1,\ldots,x_n)dx_1\ldots dx_n. \tag{45}$$

Comme pour (43), on écrit aussi $E[g(X)]=\int_{I\!R^n} g(x)f(x)dx$.

Pour simplifier, on énonce le résultat suivant pour une variable aléatoire X à valeurs dans $I\!R^2$, et on note X et Y ses deux composantes: $X=(Y,Z)$. Il se généralise sans peine à une dimension supérieure.

Proposition 17: *Supposons que X admette une densité f.*

a) Y et Z admettent les densités f_Y et f_Z suivantes sur $I\!R$:

$$f_Y(y) \;=\; \int_{-\infty}^{+\infty} f(y,z)dz, \qquad f_Z(z) \;=\; \int_{-\infty}^{+\infty} f(y,z)dy. \tag{46}$$

b) La formule suivante définit une densité sur $I\!R$, pour tout y tel que $f_Y(y)>0$:

$$f_{Z/Y=y}(z) \;=\; \frac{f(y,z)}{f_Y(y)}. \tag{47}$$

f_Y et f_Z s'appellent les **densités marginales** de f. Noter que la réciproque de (a) est fausse: les variables aléatoires Y et Z peuvent avoir des densités sans que le couple $X=(Y,Z)$ en ait une: supposons par exemple que $Z=Y$; si $\Delta=\{(x,x):x\in I\!R\}$ est la diagonale de $I\!R^2$, on a évidemment $P_X(\Delta)=1$, tandis que si la formule (45) était valide pour P_X on aurait $P_X(\Delta)=E(1_\Delta(X))=\int_{I\!R^2}1_\Delta(x)f(x)dx=0$.

L'interprétation de (47) est la suivante: la fonction $f_{Z/Y=y}$ est la densité de la "loi conditionnelle de Z sachant que $Y=y$"; bien-sûr, on a $P(Y=y)=0$ (puisque Y admet une densité), donc la phrase ci-dessus n'a pas réellement de sens, mais elle se "justifie" ainsi: Δy et Δz étant de "petits" accroissements des variables y et z, on a comme en (10), et lorsque f est continue:

$$f_Y(y)\Delta y \;\sim\; P(y \leq Y \leq y+\Delta y),$$

$$f(y,z)\Delta y\Delta z \;\sim\; P(y \leq Y \leq y+\Delta y, z \leq Z \leq z+\Delta z).$$

Par suite

$$\begin{aligned} f_{Z/Y=y}(z)\Delta z \;&\sim\; \frac{P(y \leq Y \leq y+\Delta y, z \leq Z \leq z+\Delta z)}{P(y \leq Y \leq y+\Delta y)} \\ &=\; P(z \leq Z \leq z+\Delta z / y \leq Y \leq y+\Delta y). \end{aligned}$$

Preuve de la Proposition 17. a) Pour tout $y \in \mathbb{R}$, on a par (43):

$$P(Y \leq y) \;=\; P(X \in]-\infty, y] \times \mathbb{R}) \;=\; \int_{-\infty}^{y} du \int_{-\infty}^{+\infty} dv\; f(u,v).$$

Donc si f_Y est définie par (46), on a $P(Y \leq y) = \int_{-\infty}^{y} f_Y(u)du$, ce qui montre que f_Y est la densité de Y. Pour Z on opère de la même manière.

b) Toute fonction positive d'intégrale 1 étant une densité, le résultat est évident. □

Si les composantes X_i de la variable aléatoire $X = (X_1, \ldots, X_n)$ sont dans $\mathcal{L}^2$, la **matrice des covariances** de X est la matrice $n \times n$ dont les éléments sont les $c_{i,j}$ =cov(X_i, X_j).

Proposition 18: *La matrice des covariances est symétrique non-négative.*

Preuve. La symétrie est évidente. "Non-négative" signifie que $\sum_{i,j=1} n a_i a_j c_{i,j} \geq 0$ pour tous réels a_i. Un calcul simple montre que

$$\sum a_i a_j c_{i,j} \;=\; \text{var}\left(\sum a_i X_i\right) \;\geq\; 0. \quad \square$$

Proposition 19: *Soit X un vecteur aléatoire n-dimensionnel, de matrice de covariance C. Soit A une matrice $m \times n$ et Y le vecteur aléatoire m-dimensionnel $Y = AX$. La matrice de covariance de Y est alors $C' = ACA^{\mathrm{t}}$, où A^{t} est la transposée de A.*

Preuve. Calcul immédiat. □

7 Variables aléatoires indépendantes

Commençons par reproduire, dans la situation des variables aléatoires vectorielles, les définitions 2-17 et 2-18:

Définition 20. Soit (X_n) une suite (finie ou infinie) de variables aléatoires vectorielles, i.e. chaque X_n est à valeurs dans $\mathbb{R}^{p(n)}$ pour un entier $p(n) \geq 1$.

a) Les $X_1, \ldots, X_n$ sont **indépendantes** si pour tous boréliens $A_i \in \mathcal{R}^{p(i)}$ dans les espaces correspondants $I\!R^{p(i)}$ on a

$$P(X_1 \in A_1, \ldots, X_n \in A_n) \;=\; \prod_{i=1}^{n} P(X_i \in A_i). \tag{48}$$

b) La suite infinie $(X_n)_{n \geq 1}$ est indépendante si, pour tout n fini, la famille $(X_1, \ldots, X_n)$ est indépendante.

Dans la suite, on considère seulement un couple (X, Y) de variables aléatoires. Les résultats s'étendent sans peine à une famille finie quelconque. On commence par un résultat général, dans lequel X et Y sont à valeurs dans $I\!R^m$ et $I\!R^n$ respectivement. Soit aussi g et h deux fonctions sur ces espaces, telles que $g(X)$ et $h(Y)$ soient aussi des variables aléatoires (par exemple, ce sont des fonctions continues, ou continues par morceaux).

Proposition 21: *Avec les notations précédentes, et si X et Y sont indépendantes, les variables aléatoires $g(X)$ et $h(Y)$ sont aussi indépendantes. Si de plus $g(X)$ et $h(Y)$ sont dans $\mathcal{L}^1$, alors le produit $g(X)h(Y)$ est aussi dans $\mathcal{L}^1$, et on a*

$$E[g(X)h(Y)] \;=\; E[g(X)]\; E[h(Y)]. \tag{49}$$

Preuve. La première assertion est évidente par définition même de l'indépendance. Pour le reste, on remarque d'abord que (49) se réduit à (48) si g et h sont des fonctions indicatrices. Comme les deux membres de (49) sont linéaires en g et en h, on a aussi (49) lorsque g et h sont étagées. D'après (22) on en déduit qu'on a aussi (49) pour g et h positives quelconques. Si g et h sont de signe quelconque, on a donc (49) pour $|g|$ et $|h|$, donc si $g(X)$ et $h(Y)$ sont dans $\mathcal{L}^1$ on en déduit que $g(X)h(Y) \in \mathcal{L}^1$. Enfin, dans ce cas, par différence (en considérant $g = g^+ - g^-$ et $h = h^+ - h^-$ et en développant le produit) on obtient (49) pour g et h elles-mêmes. $\square$

Si X et Y sont des variables aléatoires réelles, il découle alors de (27) et du résultat précédent que:

Proposition 22: *Si les variables aléatoires X et Y sont indépendantes et dans $\mathcal{L}^2$, on a* $\mathrm{cov}(X, Y) = 0$ *et $\rho = 0$ (ρ est le coefficient de corrélation).*

On sait que $|\rho| \leq 1$ par (29); si les variables aléatoires sont indépendantes on a donc $|\rho| = 0$; au contraire si $|\rho|$ est proche de 1, les variables aléatoires sont "fortement dépendantes", d'où le nom de "coefficient de corrélation". Cette assertion est étayée par:

Proposition 23: *Soit X et Y deux variables aléatoires dans $\mathcal{L}^2$.*

a) Si $\mathrm{var}(X) = 0$ *et si $m = E(X)$, on a $P(X = m) = 1$ (i.e. X est "presque sûrement" égale à la constante m).*

b) Si $\mathrm{var}(X) > 0$ *et* $\mathrm{var}(Y) > 0$, *on a $|\rho| = 1$ si et seulement s'il existe deux constantes $a \neq 0$ et b telle que $P(Y = aX + b) = 1$, et alors le signe de a est le même que celui de ρ.*

Preuve. a) D'après l'inégalité de Bienaymé-Tchebicheff (2-19), on a $P(|X-m| \geq \frac{1}{n}) = 0$. Comme $\{X \neq m\} = \cup_{n \geq 1}\{|X-m| \geq \frac{1}{n}\}$, on en déduit que $P(X \neq m) = 0$.

b) D'après (28), si $Y = aX+b$ on a $\text{var}(Y) = a^2\text{var}(X)$ et $\text{cov}(X,Y) = a\text{var}(X)$, donc $\rho = 1$ (resp. $= -1$) si $a > 0$ (resp. $a < 0$). Si maintenant $Y' = aX + b$ et $P(Y = Y') = 1$, on déduit de (24) que $\text{var}(Y) = \text{var}(Y')$ et $\text{cov}(X,Y) = \text{cov}(X,Y')$, de sorte que ρ prend encore les mêmes valeurs 1 ou -1.

Inversement, supposons que $|\rho| = 1$. Quitte à ajouter des constantes à X et Y, on peut supposer que $E(X) = E(Y) = 0$, ce qui ne modifie ni les variances, ni les covariances, ni ρ, ni l'existence d'une relation de type $P(Y = aX + b) = 1$ (cela modifie b, bien-sûr). D'après (28) encore, on a $\text{var}(Y - xX) = \text{var}(Y) - 2x\text{cov}(X,Y) + x^2\text{var}(X)$. Par hypothèse, ce trinôme en x admet une racine double $x = a$, et $a \neq 0$ puisque $\text{var}(Y) > 0$. On a alors $\text{var}(Y - aX) = 0$, et comme $E(Y - aX) = 0$ par hypothèse on déduit de (a) que $P(Y = aX) = 1$. $\square$

Le résultat suivant est un analogue de l'équivalence (i)$\Leftrightarrow$(ii) dans la proposition 2-12.

Proposition 24: *Soit X et Y deux variables aléatoires réelles admettant les densités f_X et f_Y. Pour qu'elles soient indépendantes, il faut et il suffit que le couple $Z = (X,Y)$ admette (sur $\mathbb{R}^2$) la densité suivante:*

$$f(x,y) \;=\; f_X(x)f_Y(y). \tag{50}$$

Preuve. Si on a indépendance, il vient par (48):

$$P(X \leq x, Y \leq y) \;=\; P(X \leq x)P(Y \leq y) \;=\; \int_{-\infty}^{x} f_X(u)du \int_{-\infty}^{y} f_Y(v)dv,$$

ce qui montre que $\mu = P_Z$ vérifie (43) avec f donnée par (50).

Inversement, supposons qu'on ait (50). Le même calcul que ci-dessus montre alors que $P(X \leq x, Y \leq y) = P(X \leq x)P(Y \leq y)$ pour tous $x, y \in \mathbb{R}$. Mais, exactemnt comme dans la preuve de la proposition 2, on peut montrer que ces relations s'étendent en $P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$ pour tous boréliens A et B, d'où le résultat. $\square$

Enfin, le problème posé à la fin du paragraphe 2-5 sur la construction des variables aléatoires de lois données se pose évidemment dans le cadre de ce chapitre. Si μ est une probabilité sur $\mathbb{R}^d$ il est de nouveau très facile de construire une variable aléatoire de loi μ: on prend $\Omega = \mathbb{R}^d$, avec pour $\mathcal{A}$ la tribu borélienne et $P = \mu$, et enfin $X(\omega) = \omega$. Pour construire une famille, même finie, de variables aléatoires de lois données il faut beaucoup plus travailler que dans le cas discret. Nous citerons donc sans démonstration l'extension suivante du théorème 2-18:

Théorème 25: *Soit, pour chaque entier n, une probabilité μ_n sur un $\mathbb{R}^{p(n)}$, où $p(n)$ est un entier non nul. Il existe alors un espace Ω muni d'une tribu $\mathcal{A}$ et d'une probabilité P, sur lequel on peut définir une suite (X_n) de variables aléatoires indépendantes, chaque X_n étant de loi μ_n.*

8 Calculs de lois

Un problème important est le suivant. Soit X une variable aléatoire réelle, admettant la densité f_X . Soit g une fonction telle que $Y = g(X)$ soit aussi une variable aléatoire. Est-ce que Y admet une densité, et si oui, comment la calculer ?

Il convient d'abord de remarquer que cette densité n'existe pas toujours. Si par exemple $g(x) = a$ pour tout x, la loi de Y est la masse de Dirac en a, qui n'a pas de densité.

Pour résoudre ce problème, l'idée consiste à essayer de mettre $E[h(Y)] = E[h \circ g(X)]$ sous la forme $\int h(y)f_Y(y)dy$ pour une fonction convenable f_Y, qui par (32) sera alors la densité cherchée (en effet (8) est un cas particulier de (32), celui où on prend $g = 1_{]-\infty,y]}$). Or, (32) implique

$$E[h(Y)] \;=\; E[h \circ g(X)] \;=\; \int_{-\infty}^{+\infty} h \circ g(x) f_X(x)dx, \tag{51}$$

et on fait le changement de variable $y = g(x)$ dans cette intégrale. Cela nécessite que g soit dérivable et bijective "par morceaux", et il faut faire très attention aux domaines où g est croissante ou décroissante. Plutôt qu'exposer une théorie générale, donnons des exemples.

Exemple:

1) Soit $X = aX + b$, où a et b sont des constantes. Si $a = 0$, on a alors $Y = b$ et la loi de Y (sans densité) est la masse de Dirac en b. Si au contraire $a \neq 0$, on fait le changement de variable $y = ax + b$ dans (51), ce qui donne

$$E[h(Y)] \;=\; \int_{-\infty}^{+\infty} h(ax+b)f_X(x)dx \;=\; \int_{-\infty}^{+\infty} h(y) f_X\left(\frac{y-b}{a}\right)\frac{1}{|a|}dy$$

(on peut considérer séparément les cas $a > 0$ et $a < 0$). Donc

$$f_Y(y) \;=\; f_X\left(\frac{y-b}{a}\right)\frac{1}{|a|}. \tag{52}$$

Par exemple:

- Si X suit la loi normale $\mathcal{N}(m, \sigma^2)$, alors $\frac{X-m}{\sigma}$ suit la loi $\mathcal{N}(0, 1)$.
- Si X suit la loi normale $\mathcal{N}(0, 1)$, alors $aX + b$ suit la loi $\mathcal{N}(b, a^2)$.
- Si X suit la loi uniforme sur $[\alpha, \beta]$, alors $aX + b$ suit la loi uniforme sur $[a\alpha + b, a\beta + b]$.
- Si X suit la loi $\Gamma(\alpha, \theta)$, alors aX suit la loi $\Gamma(\alpha, \theta/a)$.

2) Soit $Y = X^2$. La fonction g est décroissante sur $I\!R_-$ et croissante sur $I\!R_+$. Le changement de variable $y = x^2$ donne alors

$$\begin{aligned} E[h(Y)] &= \textstyle\int_{-\infty}^{0} h(x^2)f_X(x)dx + \int_0^{+\infty} h(x^2)f_X(x)dx \\ &= \textstyle\int_0^{+\infty} h(y)f_X(-\sqrt{y})\frac{1}{2\sqrt{y}}dy + \int_0^{+\infty} h(y)f_X(\sqrt{y})\frac{1}{2\sqrt{y}}dy, \end{aligned}$$

et on a donc

$$f_Y(y) \;=\; (f_X(-\sqrt{y}) + f_X(\sqrt{y}))\;\frac{1}{2\sqrt{y}}. \tag{53}$$

Par exemple, si X suit la loi $\mathcal{N}(0, 1)$, X^2 suit la loi $\Gamma(1/2, 1/2)$.

Dans le cas des vecteurs aléatoires, l'idée est la même. Soit $X = (X_1, \ldots, X_n)$ une variable aléatoire de densité f_X sur $\mathbb{R}^n$. Soit g une fonction de $\mathbb{R}^n$ dans $\mathbb{R}^m$, et $Y = g(X)$. Plusieurs cas sont à considérer:

a) On a $m > n$**:** Le vecteur Y n'admet pas de densité.

b) On a $m = n$**:** On fait, comme dans le cas unidimensionnel, le changement de variable $y = g(x)$ dans l'intégrale n-uple qui remplace le terme de droite de (51).

Rappelons cette formule de changement de variable. D'abord, supposons que g soit une bijection continûment différentiable de A dans B (deux ouverts de $\mathbb{R}^n$). Son **jacobien** est le déterminant $J(x)$ dont les composantes sont les dérivéees partielles $\partial g_i(x)/\partial x_j$, où g_i est la $i^{\text{ème}}$ composante de g. On a alors

$$\int_A h \circ g(x) f_X(x) dx \ = \ \int_B h(y) f_X \circ g^{-1}(y) \frac{1}{|J(g^{-1}(y))|} dy \tag{54}$$

(attention à la valeur absolue de J). Rappelons d'ailleurs que $1/J(g^{-1}(y))$ est le jacobien de la transformation inverse g^{-1} au point $y \in B$. Si alors $f_X(x) = 0$ en dehors de A, on obtient que Y admet la densité

$$f_Y(y) \ = \ 1_B(y) f_X \circ g^{-1}(y) \frac{1}{|J(g^{-1}(y))|}. \tag{55}$$

Lorsque g est simplement continûment différentiable, il existe souvent une partition finie $(A_i)_{1 \le i \le d}$ de l'ensemble $\{x : f_X(x) > 0\}$, telle que g soit injective sur chaque A_i, et on note $B_i = g(A_i)$ l'image de A_i par g. On découpe alors l'intégrale selon les A_i, on applique (54) à chaque morceau, et on fait la somme. On obtient alors

$$f_Y(y) \ = \ \sum_{i=1}^{d} 1_{B_i}(y) f_X \circ g^{-1}(y) \frac{1}{|J(g^{-1}(y))|}, \tag{56}$$

où g^{-1} est bien définie sur chaque B_i (comme image réciproque de la restriction de g à A_i).

On a $m < n$**:** On commence par "compléter" Y, en essayant de construire une application g' de $\mathbb{R}^n$ dans $\mathbb{R}^n$ dont les m premières composantes coïncident avec les composantes de g, et pour laquelle on puisse appliquer (55) ou (56). On obtient ainsi la densité $f_{Y'}$ de $Y' = g'(X)$. Puis on applique l'extension évidente de (46):

$$f_Y(y_1, \ldots, y_m) \ = \ \int \ldots \int_{\mathbb{R}^{n-m}} f_{Y'}(y_1, \ldots, y_m, y_{m+1}, \ldots, y_n) dy_{m+1} \ldots dy_n. \tag{57}$$

Exemples.

1) Coordonnées polaires: soit $X = (U, V)$ un vecteur aléatoire de $\mathbb{R}^2$, et $Y = (R, \Theta)$ ses coordonnées polaires. La transformation g est bijective de $A = \mathbb{R}^2 \backslash \{0\}$ dans $B =]0, \infty[\times]0, 2\pi]$, et son inverse g^{-1} s'écrit facilement: $u = r \cos\theta$, $v = r \sin\theta$. Le jacobien de g^{-1} au point (r, θ) est r, donc (55) entraine

$$f_Y(r, \theta) \ = \ r \ f_X(r\cos\theta, r\sin\theta) 1_B(r, \theta). \tag{58}$$

Par exemple si U et V sont indépendantes et de loi $\mathcal{N}(0,1)$, (50) entraine que $f_X(u,v) = \frac{1}{2\pi} \exp -\frac{u^2+v^2}{2}$, donc (58) implique

$$f_Y(r,\theta) = \frac{1}{2\pi} r\, e^{-r^2/2} 1_{]0,\infty[}(r) 1_{]0,2\pi]}(\theta). \tag{59}$$

En particulier les variables aléatoires R et Θ sont indépendantes, la première suit la loi de densité $re^{-r^2/2}1_{]0,\infty[}(r)$, et la seconde est uniforme sur $[0,2\pi]$.

2) Soit $X = (U,V)$ un vecteur aléatoire de $\mathbb{R}^2$, avec U et V indépendantes de lois $\Gamma(\alpha,\theta)$ et $\Gamma(\beta,\theta)$. Quelle est la densité de $Y = \frac{U}{U+V}$?

Comme la dimension de Y est plus petite que celle de X, il faut d'abord "compléter" Y. On prend par exemple $Y' = (Y,Z)$, avec $Z = U+V$, ce qui correspond à $g(u,v) = \left(\frac{u}{u+v}, u+v\right)$. Cette application est bijective de $A =]0,\infty[^2$ dans $B =]0,1[\times]0,\infty[$, et on a $g^{-1}(y,zy) = (yz, z(1-y))$, qui a pour jacobien z. Comme $f_X(u,v) = \frac{\theta^{\alpha+\beta}}{\Gamma(\alpha)\Gamma(\beta)} u^{\alpha-1} v^{\beta-1} e^{-\theta(u+v)} 1_A(u,v)$, (55) entraine

$$f_{Y'}(y,z) = \frac{\theta^{\alpha+\beta}}{\Gamma(\alpha)\Gamma(\beta)} z^{\alpha+\beta-1} y^{\alpha-1}(1-y)^{\beta-1} e^{-\theta z} 1_B(y,z). \tag{60}$$

Il reste à appliquer (57):

$$\begin{aligned} f_Y(y) &= \int f_{Y'}(y,z)dz \\ &= \frac{\theta^{\alpha+\beta}}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1} 1_{]0,1[}(y) \int_0^\infty z^{\alpha+\beta-1} e^{-\theta z} dz \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1} 1_{]0,1[}(y) \end{aligned} \tag{61}$$

(utiliser (36)). On appelle **loi béta** de paramètres α et β la loi admettant cette densité.

On obtient aussi facilement la densité de Z: en effet, (61) montre que $f_{Y'}(y,z)$ est le produit de $f_Y(y)$ par la fonction

$$f_Z(z) = \frac{\theta^{\alpha+\beta}}{\Gamma(\alpha)\Gamma(\beta)} z^{\alpha+\beta-1} e^{-\theta z} 1_{]0,\infty[}(z),$$

qui d'après (46) est la densité de la variable aléatoire Z. On a en fait démontré la:

Proposition 26: *Si U et V sont indépendantes et de lois respectives $\Gamma(\alpha,\theta)$ et $\Gamma(\beta,\theta)$ alors $U+V$ suit la loi $\Gamma(\alpha+\beta,\theta)$ et est indépendante de $\frac{U}{U+V}$.*

3) Si U et V sont indépendantes de densités f_U et f_V, on peut de la même manière trouver la densité de la somme $U+V$. Là encore on "complète" en $T = (U,Z)$ (par exemple), correspondant à la bijection $g(u,v) = (u, u+v)$ sur $\mathbb{R}^2$, dont le jacobien est 1. Par suite (55) conduit à:

Proposition 27: *Si U et V sont deux variables aléatoires indépendantes de densités f_U et f_V, alors $Z = U + V$ admet la densité*

$$f_Z(z) \;=\; \int_{-\infty}^{+\infty} f_U(u) f_V(z-u)du \;=\; \int_{-\infty}^{+\infty} f_U(z-v) f_V(v)dv. \qquad (62)$$

Etant donnée l'importance de la formule ci-dessus, on lui donne un nom: on dit que la fonction f_Z est le **produit de convolution** des deux fonctions f_U et f_V. Nous verrons intervenir de nouveau le produit de convolution dans le chapitre suivant.

9 Simulation de variables aléatoires

Une question naturelle était, nous l'avons vu, de construire une variable aléatoire de loi donnée. La "suite" de cette question, tout aussi naturelle sur le plan des applications, consiste à "simuler" cette variable aléatoire: cela permet, combiné avec la loi des grands nombres que nous verrons au chapitre suivant, de faire beaucoup de calculs numériques (calculs d'intégrales notamment). Bien entendu, il convient pour cela de simuler un grand nombre de variables aléatoires de loi donnée, et donc le domaine de la simulation n'a pu se développer de manière significative que depuis l'introduction massive des ordinateurs.

Nous allons ci-dessous présenter deux méthodes simples, mais relativement générales, pour faire une telle simulation. Il existe, pour un certain nombre de lois courantes (la loi normale notamment) des méthodes ad-hoc que nous n'expliquons pas ici.

Le principe de base est ce qu'on appelle un **générateur de nombres au hasard**. C'est un algorithme qui fournit une suite de nombres compris entre 0 et 1, qui ont les mêmes caractéristiques statistiques qu'une suite de variables aléatoires indépendantes et de loi uniforme sur $[0,1]$. C'est typiquement ce que fournit une application répétée de la fonction "random" dans la plupart des ordinateurs. Bien entendu, la qualité de ces générateurs est fondamentale, mais nous supposerons ce problème résolu de manière correcte: en d'autres termes, cela signifie qu'on dispose d'une suite $X_1, \ldots$ (potentiellement infinie) de variables aléatoires indépendantes, uniformes sur $[0,1]$. Nous visons à construire une suite $Y_1, \ldots$ de variables aléatoires réelles, de loi μ fixée.

1) Inversion de la fonction de répartition: La première méthode consiste à utiliser la fonction de répartition F de μ, et son "inverse" défini ainsi:

$$G(x) \;=\; \inf(y : F(y) > x) \qquad \forall x \in]0,1[. \qquad (63)$$

La fonction G n'est pas à proprement parler l'inverse (ou fonction réciproque) de F, puisque celle-ci n'est pas nécessairement bijective de $\mathbb{R}$ dans $]0,1[$. Elle joue cependant le même rôle, dans la mesure où elle vérifie $F(G(x)) = x$ si $0 < x < 1$ et si la fonction F est continue au point $G(x)$.

Proposition 28: *Si on pose $Y_n = G(X_n)$, on obtient une suite de variables aléatoires indépendantes, de loi μ.*

Preuve. L'indépendance et le fait que les Y_n suivent la même loi sont évidents. On remarque facilement que si $x \in \mathbb{R}$ et $0 < y < 1$, on a $G(y) < x \;\Rightarrow\; y < F(x) \;\Rightarrow\; G(y) \leq x$.

Par suite

$$P(Y_n < x) = P(G(X_n) < x) \leq P(X_n < F(x)) = F(x) \leq P(G(X_n) \leq x) = P(Y_n \leq x)$$

puisque X_n suit une loi uniforme sur $[0,1]$. Comme $x \mapsto P(Y_n \leq x)$ est continue à droite et croissante, on en déduit que $P(Y_n \leq x) = F(x)$ pour tout x, donc Y_n suit la loi μ. $\square$

2) Méthode du rejet: Cette méthode s'applique lorsque la probabilité μ admet une densité f, et lorsqu'on connait une autre probabilité ν, selon laquelle on peut simuler des variables aléatoires et qui admet une densité g telle que

$$f(x) \ \leq \ ag(x), \qquad \forall x \in I\!R^d, \tag{64}$$

pour une constante connue a (nécessairement $a \geq 1$, et même $a > 1$ si $\mu \neq \nu$, puisque f et g sont deux fonctions positives ayant la même intégrale 1).

On suppose aussi qu'on dispose, en sus de la suite $X_1, \ldots$ ci-dessus (constituée de variables aléatoires indépendantes de loi uniforme), d'une suite $Z_1, \ldots$ potentiellement infinie de variables aléatoires indépendantes de loi ν et aussi indépendantes des X_n. On pose alors

$$N \ = \ \inf(n \in I\!N^\star : f(Z_n) > aX_n g(Z_n)), \qquad Y = Z_n \quad \text{si} \quad N = n. \tag{65}$$

Proposition 29: *La variable aléatoire N est à valeurs dans $I\!N^\star$, et la variable aléatoire Y suit la loi μ.*

Preuve. La première assertion entraîne que la variable Y est bien définie dans la seconde formule (65). Notons $A_n = \{f(Z_n) > aX_n g(X_n)\}$. Les événements A_n sont indépendants, et $P(A_n) = \alpha$ (on calculera α plus tard), de sorte que $P(N > n) = (1-\alpha)^n$, et pour toute fonction h il vient

$$E[h(Y)] \ = \ \sum_{n=1}^{\infty} E[h(Z_n) 1_{A_n} 1_{\{N > n-1\}}].$$

Les variables aléatoires $h(Z_n)1_{A_n}$ d'une part et $1_{\{N>n-1\}}$ d'autre part sont indépendantes, donc

$$E[h(Y)] \ = \ \sum_{n=1}^{\infty} E[h(Z_n) 1_{A_n}](1-\alpha)^{n-1}. \tag{66}$$

Enfin $E[h(Z_n)1_{A_n}]$ vaut

$$\int h(z)g(z)dz \int_0^1 1_{\{f(z) > axg(z)\}} dx = \int h(z)g(z) \frac{f(z)}{ag(z)} dz = \frac{1}{a} \int h(z)f(z)dz = \frac{1}{a} \int h d\mu.$$

En particulier α égale cette expression lorsque $h = 1$, donc $\alpha = 1/a$. En remplaçant dans (66), on obtient que $E[h(Y)] = \int h d\mu$, d'où le résultat. $\square$

Cela donne une variable aléatoire de loi μ. Pour obtenir une suite de telles variables aléatoires indépendantes, il faut répéter la même procédure.

On peut comparer les deux méthodes. La première est très simple à mettre en oeuvre, **si on connait explicitement la fonction** G, ce qui est assez rare dans la pratique. La seconde nécessite la connaissance de f, g et a, et aussi le fait qu'on sache préalablement simuler selon la loi ν (si on peut par exemple utiliser la première méthode pour cette loi): ces conditions sont assez souvent remplies. Elle est malheureusement parfois longue à mettre en oeuvre (sa "longueur" est proportionnelle à N, qui est aussi une variable aléatoire: on ne contrôle donc pas très bien la longueur de la procédure).

CHAPITRE 4

Fonctions caractéristiques et théorèmes limites

1 La fonction caractéristique

Dans ce paragraphe nous introduisons un outil important en calcul des probabilités: il s'agit de ce qu'on appelle la fonction caractéristique d'une variable aléatoire, et qui dans d'autres branches des mathématiques s'appelle aussi **transformée de Fourier**.

On notera $< x, y >$ le produit scalaire de deux vecteurs de $\mathbb{R}^n$. Si $u \in \mathbb{R}^n$, la fonction (complexe) $x \mapsto e^{i<u,x>}$ est continue, de module 1. Donc si X est une variable aléatoire à valeurs dans $\mathbb{R}^n$, on peut considérer $e^{i<u,X>}$ comme une variable aléatoire à valeurs complexes (i.e., ses parties réelle $Y = \cos(< u, X >)$ et imaginaire $Z = \sin(< u, X >)$ sont des variables aléatoires réelles). Ces variables aléatoires réelles sont en plus bornées par 1, donc elles admettent une espérance. Il est alors naturel d'écrire que l'espérance de $e^{i<u,X>}$ est $E(e^{i<u,X>}) = E(Y) + iE(Z) = E(\cos(< u, X >)) + iE(\sin(< u, X >))$.

Définition 1. Si X est une variable à valeurs dans $\mathbb{R}^n$, sa **fonction caractéristique** est la fonction ϕ_X de $\mathbb{R}^n$ dans $\mathbb{C}$ définie par

$$\phi_X(u) \;=\; E\left(e^{i<u,X>}\right). \tag{1}$$

On remarquera que la fonction caractéristique ne dépend en fait **que de la loi** P_X de X. C'est en fait la "transformée de Fourier" de la loi P_X.

Pour pouvoir démontrer les propriétés de ces fonctions, nous aurons besoin d'un résultat très important de théorie de l'intégration (c'est ce résultat qui, en grande partie, fait la supériorité de l'intégrale de Lebesgue par rapport à celle de Riemann). Bien qu'il ne soit pas très difficile, il s'appuie sur un certain nombre de propriétés de l'espérance que nous n'avons pas vues, et nous l'énonçons donc sans démonstration:

Théorème 2 (de Lebesgue, ou de convergence dominée): *Si les variables aléatoires réelles ou complexes Y_n convergent simplement sur Ω vers une limite Y et si $|Y_n| \leq Z$*

pour tout n, où $Z \in \mathcal{L}^1$, alors Y_n et Y sont dans $\mathcal{L}^1$ et on a $E(Y_n) \to E(Y)$.

Proposition 3: *ϕ_X est une fonction de module inférieur ou égal à 1, continue, avec $\phi_X(0) = 1$.*

Preuve. La dernière assertion est évidente. Comme $E(Y)^2 \leq E(Y^2)$ pour toute variable aléatoire réelle Y, il vient (pour un nombre complexe z, $|z|$ désigne le module):

$$|\phi_X(u)|^2 \;=\; E(\cos < u, X >)^2 + E(\sin < u, X >)^2 \;\leq\; E(\cos^2 < u, X > + \sin^2 < u, X >),$$

et donc $|\phi_X(u)| \leq 1$.

Enfin, pour la continuité, si $u_p \to u$, on a convergence simple des $e^{i<u_p,X>}$ vers $e^{i<u,X>}$ et la majoration du module par la constante 1, qui est dans $\mathcal{L}^1$: par suite $\phi_X(u_p) \to \phi_X(u)$, et la fonction ϕ_X est continue. $\square$

Proposition 4: *Si la variable aléatoire $|X|^m$ (où $|X|$ désigne la norme euclidienne du vecteur X) est dans $\mathcal{L}^1$ pour un entier m, la fonction ϕ_X est m fois continûment différentiable sur $I\!R^n$, et on a pour tout choix des indices $i_1, \ldots, i_m$:*

$$\frac{\partial^m}{\partial x_{i_1} \partial x_{i_2} \ldots \partial x_{i_m}} \phi_X(u) \;=\; i^m \; E\left(e^{i<u,X>} \; X_{i_1} X_{i_2} \ldots X_{i_m}\right) \tag{2}$$

(les X_j sont les composantes de X).

En prenant $u = 0$ ci-dessus, cette formule permet de calculer $E\left(X_{i_1} X_{i_2} \ldots X_{i_m}\right)$ en fonctions des dérivées à l'origine de ϕ_X. Par exemple, si X est 1-dimensionnelle, on a

$$E(X) = i\; \phi'_X(0), \qquad (\text{resp.} \quad E(X^2) = -\phi''_X(0)\ ) \tag{3}$$

dès que $X \in \mathcal{L}^1$ (resp. $X \in \mathcal{L}^2$).

Preuve. On se contente du cas $m = 1$, le cas général se montrant de la même manière, par récurrence sur m. Soit v le $j^{\text{ème}}$ vecteur de base de $I\!R^n$. On a

$$\frac{\phi_X(u + tv) - \phi_X(u)}{t} \;=\; E\left(e^{i<u,X>} \; \frac{e^{itX_j} - 1}{t}\right). \tag{4}$$

Soit $t_p \to 0$. Les variables aléatoires $(e^{it_p X_j} - 1)/t_p$ convergent simplement vers iX_j, en restant bornées en module par la variable aléatoire $2|X_j|$, qui par hypothèse est dans $\mathcal{L}^1$. Donc par le théorème 2 (de Lebesgue) on en déduit que (4) converge vers $iE(e^{i<u,X>}\; X_j)$ quand $t \to 0$. On en déduit que la première dérivée partielle de ϕ_X par rapport à x_j existe et est donnée par la formule (2). Enfin, on montre comme dans la proposition précédente que cette dérivée est continue. $\square$

Proposition 5: *Si X est une variable aléatoire à valeurs dans $I\!R^n$, si $a \in I\!R^m$ et si A est une matrice $m \times n$, on a:*

$$\phi_{a+AX}(u) \;=\; e^{i<a,u>}\; \phi_X(A^{\mathrm{t}} u), \qquad \forall u \in I\!R^m. \tag{5}$$

Preuve. On a $e^{i<u,a+AX>} = e^{i<u,a>}\ e^{i<A^{\mathrm{t}},X>}$, et il suffit de prendre les espérances pour obtenir le résultat. $\square$

Exemple:

1) Loi binomiale $B(p,n)$: $\phi(u) = (pe^{iu} + 1 - p)^n$.

2) Loi de Poisson de paramètre θ: $\phi(u) = e^{\theta(e^{iu}-1)}$.

3) Loi uniforme sur $[a,b]$: $\phi(u) = \frac{\sin(au)}{au}$.

4) Loi normale $\mathcal{N}(0,1)$: en intégrant la fonction de variable complexe $\frac{1}{\sqrt{2\pi}} \exp -z^2/2$ sur le contour ci-dessous, puis en faisant $a \uparrow +\infty$,

$-a$ 0 a

$-iu$

on arrive à:

$$\phi(u) \;=\; e^{-u^2/2}. \tag{6}$$

5) Loi normale $\mathcal{N}(m,\sigma^2)$: on sait qu'une variable aléatoire X admettant cette loi s'écrit $X = m + \sigma Y$, où Y suit la loi $\mathcal{N}(0,1)$. D'après (5) et (6) on a donc

$$\phi(u) \;=\; e^{ium - u^2\sigma^2/2}. \tag{7}$$

6) Loi exponentielle de paramètre θ: $\phi(u) = \frac{\theta}{\theta - iu}$.

7) Loi gamma $\Gamma(\alpha,\theta)$: en intégrant la fonction de variable complexe $z^{\alpha-1}\ e^{-\theta z}$ sur le contour ci-dessous, puis en faisant $c \downarrow 0$ et $d \uparrow +\infty$,

0 c d

droite d'équation $y = -ux/a$

on arrive à:

$$\phi(u) \;=\; \frac{\theta^\alpha}{(\theta - iu)^\alpha}. \tag{8}$$

L'un des intérêts majeurs des fonctions caractéristiques réside dans la propriété suivante (d'où provient aussi leur nom):

Théorème 6: *La fonction caractéristique ϕ_X caractérise la loi de la variable aléatoire X.*

Preuve. Soit les fonctions suivantes sur $I\!R^n$, avec $\sigma > 0$:

$$f_\sigma(x) \;=\; \frac{1}{(2\pi\sigma^2)^{n/2}}\; e^{-|x|^2/2\sigma^2}, \qquad \widehat{f}_\sigma(u) \;=\; e^{-|u|^2\sigma^2/2}.$$

On a

$$\begin{aligned}\int f_\sigma(x)e^{i<u,x>}ds \;&=\; \int \prod_{j=1}^{n}\left(\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{x_j^2}{2\sigma^2}+iu_jx_j\right)\right)dx_1\dots dx_n\\ &=\; \prod_{j=1}^{n}\int \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{t^2}{2\sigma^2}+iu_jt\right)dt \;=\; \widehat{f}_\sigma(u)\end{aligned}$$

d'après l'exemple 4. Donc

$$f_\sigma(u-v) \;=\; \frac{1}{(2\pi\sigma^2)^{n/2}}\widehat{f}_\sigma\left(\frac{u-v}{\sigma^2}\right) \;=\; \frac{1}{(2\pi\sigma^2)^{n/2}}\int f_\sigma(x)e^{i<u-v,x>/\sigma^2}dx.$$

Supposons alors que X et X' admettent la même fonction caractéristique $\phi_X = \phi_{X'}$. Alors, en admettant qu'on puisse échanger l'intégration par rapport à "dx" (sur $I\!R^n$) et l'espérance, il vient

$$\begin{aligned}E[f_\sigma(X-v)] \;&=\; E\left(\frac{1}{(2\pi\sigma^2)^{n/2}}\int f_\sigma(x)e^{i<X-v,x>/\sigma^2}dx\right)\\ &=\; \int f_\sigma(x)\;\frac{1}{(2\pi\sigma^2)^{n/2}}\;\phi_X\left(\frac{x}{\sigma^2}\right)e^{-i<v,x>/\sigma^2}dx,\end{aligned} \tag{9}$$

et de même pour X'. Par suite $E[f(X)] = E[f(X')]$ pour toute fonction de la forme $f(u) = f_\sigma(u-v)$ pour $v \in I\!R$ et $\sigma > 0$ arbitraires, donc aussi pour toute fonction f dans l'espace vectoriel engendré par ces fonctions. D'après le théorème de Stone-Weierstass, cet espace est dense dans l'ensemble C_0 des fonctions continues sur $I\!R^n$ et ayant une limite nulle à l'infini, pour la topologie de la convergence uniforme. Par suite $E[f(X)] = E[f(X')]$ pour toute $f \in C_0$. Comme l'indicatrice d'un ouvert quelconque est limite croissante de fonctions de C_0, on en déduit que $P_X(A) = E[1_A(X)]$ égale $P_{X'}(A) = E[1_A(X')]$ pour tout ouvert A, ce qui implique $P_X = P_{X'}$.

Corollaire 7: *Soit $X = (X_1,\dots,X_n)$ une variable aléatoire à valeurs dans $I\!R^n$. Pour que les composantes X_i soient indépendantes, il faut et il suffit que pour tous $u_1,\dots,u_n \in I\!R$ on ait*

$$\phi_X(u_1,\dots,u_n) \;=\; \prod_{j=1}^{n}\;\phi_{X_j}(u_j). \tag{10}$$

Preuve. Comme $e^{<u,X>} = \prod_j e^{i<u_j,X_j>}$, la nécessité découle de (3-49). Supposons inversement qu'on ait (10). D'après le théorème 3-25 on peut construire des variables aléatoires X_j'

indépendantes, telles que X'_j et X_j aient mêmes lois pour tout j, donc $\phi_{X'_j} = \phi_{X_j}$. Si $X' = (X'_1, \ldots, X'_n)$ on a donc $\phi_{X'} = \phi_X$ par (10) et la condition nécessaire. Donc X et X' ont même loi, ce qui entraine que pour tous boréliens A_j on ait

$$P(\cap_j \{X_j \in A_j\}) \;=\; P((\cap_j \{X'_j \in A_j\}) \;=\; \prod_j P(X'_j \in A_j) \;=\; \prod_j P(X_j \in A_j),$$

d'où l'indépendance cherchée. □

La transformée de Laplace: Lorsque X est un variable aléatoire à valeurs dans $\mathbb{R}_+$, on définit sa transformée de Laplace par

$$\psi_X(\lambda) \;=\; E\left(e^{-\lambda X}\right), \qquad \lambda \in \mathbb{R}_+. \tag{11}$$

C'est une fonction définie sur $\mathbb{R}_+$, indéfiniment dérivable sur $]0,\infty[$, et formellement on a $\psi_X(\lambda) = \phi_X(i\lambda)$; ainsi, il n'est pas étonnant que la transformée de Laplace ait des propriétés analogues à celles de la fonction caractéristique. En particulier, elle caractérise la loi P_X.

Si de plus X est une variable aléatoire à valeurs dans $\mathbb{N}$, de fonction génératrice g_X, alors on a $\psi_X(\lambda) = g_X(e^{-\lambda})$.

2 Somme de variables aléatoires indépendantes

Vu ce qui précède, on généralise naturellement la proposition 2-16 par:

Proposition 8: *Si X et Y sont deux variables aléatoires indépendantes à valeurs dans $\mathbb{R}^n$, la fonction caractéristique de la somme $X+Y$ est donnée par*

$$\phi_{X+Y} \;=\; \phi_X \phi_Y. \tag{12}$$

Preuve. Comme $e^{i<u,X+Y>} = e^{i<u,X>}e^{i<u,Y>}$, il suffit d'appliquer (3-49). □

Proposition 9: *Si les X_j sont des variables aléatoires indépendantes et dans $\mathcal{L}^2$, les variances vérifient*

$$\mathrm{var}\left(\sum_{i=1}^n X_i\right) \;=\; \sum_{i=1}^n \mathrm{var}(X_i). \tag{13}$$

Preuve. D'après la linéarité de l'espérance et la définition de la covariance, il est clair que

$$\mathrm{var}\left(\sum_{i=1}^n X_i\right) \;=\; \sum_{i,j=1}^n \mathrm{cov}(X_i, X_j).$$

Comme ici $\mathrm{cov}(X_i, X_j) = 0$ si $i \neq j$, on a le résultat. □

Exemple. Soit X, Y indépendantes, et $Z = X + Y$:

1) X suit $\Gamma(\alpha,\theta)$ et Y suit $\Gamma(\beta,\theta)$: alors Z suit $\Gamma(\alpha+\beta,\theta)$ (c'est la proposition 3-26; cela découle aussi de (8) et (12)).

2) X suit $\mathcal{N}(m,\sigma^2)$ et Y suit $\mathcal{N}(m',\sigma'^2)$: alors Z suit $\mathcal{N}(m+m',\sigma^2+\sigma'^2)$ (cela découle aussi de (7) et (12)).

3) X et Y suivent des lois de Poisson de paramètres θ et θ': alors Z suit une loi de Poisson de paramètre $\theta+\theta'$ (encore (12)).

4) X suit $B(p,n)$ et Y suit $B(p,m)$: alors Z suit $B(p,n+m)$) (encore (12)).

3 Vecteurs gaussiens

Définition 10. Une variable aléatoire $X=(X_1,\dots,X_n)$ à valeurs dans $\mathbb{R}^n$ est appelée un **vecteur gaussien** si toute combinaison linéaire $\sum_{j=1}^n a_jX_j$ suit une loi normale (avec la convention que la masse de Dirac au point m est la "loi normale" $\mathcal{N}(m,0)$).

Cela entraîne bien entendu que chaque composante X_j suit elle-même une loi normale.

Exemple: Si les X_i sont des variables aléatoires normales indépendantes, le vecteur X est gaussien (utiliser l'exemple 2 du paragraphe 2).

Contre-exemple: Si les composantes X_i sont normales mais pas indépendantes, il se peut que X ne soit pas gaussien. Prenons par exemple X_1 de loi $\mathcal{N}(0,1)$, et

$$X_2 \;=\; \begin{cases} X_1 & \text{si } \; |X_1|\leq 1 \\ -X_1 & \text{sinon.} \end{cases}$$

Alors X_2 suit également la loi $\mathcal{N}(0,1)$, mais $X=(X_1,X_2)$ n'est pas un vecteur gaussien, puisque $0<P(X_1+X_2=0)<1$ (donc X_1+X_2 ne suit pas une loi normale).

Théorème 11: *X est un vecteur gaussien si et seulement si sa fonction caractéristique s'écrit*

$$\phi_X(u) \;=\; e^{i<u,m>-\frac{1}{2}<u,Cu>}, \tag{14}$$

où $m\in\mathbb{R}^n$ et C est une matrice symétrique $n\times n$ non-négative; dans ce cas on a $m=E(X)$ (i.e. $m_j=E(X_j)$ pour chaque j) et C est la matrice des covariances de X.

Preuve. a) Condition suffisante: supposons (14). Pour toute combinaison linéaire $Y=\sum_j a_jX_j=<a,X>$ on a, pour $v\in\mathbb{R}$:

$$\phi_Y(v) \;=\; \phi_X(va) \;=\; e^{iv<m,a>-\frac{v^2}{2}<a,Ca>},$$

donc Y suit la loi $\mathcal{N}(<a,m>,<a,Ca>)$.

b) Condition nécessaire: Soit C la matrice des covariances de X et m son vecteur moyenne (noter que ces quantités existent, car chaque composante X_j, étant gaussienne, est dans $\mathcal{L}^2$). Si $Y=<a,X>$ avec $a\in\mathbb{R}^n$, on a

$$E(Y) \;=\; <a,m>, \qquad \text{var}(Y) \;=\; <a,Ca>.$$

Par hypothèse Y suit une loi normale, donc vu ce qui précède sa fonction caractéristique est

$$\phi_Y(v) \;=\; e^{iv<a,m>-\frac{v^2}{2}<a,Ca>}.$$

Mais $\phi_Y(1) = \phi_{<a,X>}(1) = E(e^{i<a,X>}) = \phi_X(a)$, d'où (14). $\square$

Corollaire 12: *Si X est un vecteur gaussien, ses composantes sont indépendantes si et seulement si la matrice de covariance est diagonale.*

Attention: ce résultat peut être faux si X n'est pas gaussien !

Preuve. Il suffit de combiner (14) et le corollaire 7. $\square$

Proposition 13: *Soit X un vecteur gaussien de moyenne m. Il existe des variables aléatoires réelles indépendantes $Y_1, \dots, Y_n$ de lois normales $\mathcal{N}(0, \lambda_j)$ avec $\lambda_j \geq 0$ (si $\lambda_j = 0$ on convient que $Y_j = 0$) et une matrice orthogonale A telles que $X = m + AY$, où $Y = (Y_1, \dots Y_n)$.*

Preuve. Comme C est une matrice symétrique non-négative (cf. proposition 3-18), il existe une matrice orthogonale A et une matrice diagonale Λ dont les éléments diagonaux vérifient $\lambda_j \geq 0$, et telle que la matrice de covariance de X s'écrive $C = A\Lambda A^{\mathrm{t}}$. Soit $Y = A^{\mathrm{t}}(X - m)$ qui est un vecteur gaussien de covariance $C' = A^{\mathrm{t}}CA = \Lambda$ et de moyenne nulle. Les composante Y_j de Y répondent à la question. $\square$

Corollaire 14: *Le vecteur gaussien X admet une densité sur $\mathbb{R}^n$ si et seulement si sa matrice de covariance est non-dégénérée (ou inversible, ou de valeurs propres toutes strictement positives).*

Preuve. Reprenons la preuve de la proposition précédente: les λ_j qui y paraissent sont les valeurs propres de C.

Si $\lambda_j > 0$ pour tout j, le vecteur aléatoire Y admet la densité suivante sur $\mathbb{R}^n$:

$$f_Y(y) \;=\; \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi\lambda_j}}\, e^{-y_j^2/2\lambda_j}.$$

Comme $X = m + AY$ on en déduit que X admet la densité

$$f_X(x) \;=\; \frac{1}{(2\pi)^{n/2}\sqrt{\text{dét}(C)}}\, e^{-\frac{1}{2}<x-m,C^{-1}(x-m>}. \tag{15}$$

Si au contraire C n'est pas inversible, il existe $a \in \mathbb{R}^n$ tel que $a \neq 0$ et $Ca = 0$. La variable aléatoire $Z =< X, a >$ a pour variance $< a, Ca >= 0$ et pour moyenne $z = < m, a >$, donc $P(Z = z) = 1$ par la proposition 3-23. Ainsi, avec une probabilité 1, le vecteur X est dans un hyperplan H orthogonal à a, i.e. $P(X \in H) = 1$. Or, si X admettait la densité f, on aurait $P(X \in H) = \int_H f(x)dx$ par (3-45) et (3-18), donc $P(X \in H) = 0$ pusique le "volume" de l'hyperplan H est nul. $\square$

4 Convergence en loi

On considère maintenant des variables aléatoires X_n et X, toutes à valeurs dans le même espace $I\!R^d$, mais pouvant éventuellement être définies sur des espace d'états différents.

Définition 15. On dit que la suite (X_n) **converge en loi** vers X, et on écrit $X_n \stackrel{\mathcal{L}}{\to} X$, si pour toute fonction f dans l'espace $C(I\!R^d)$ des fonctions continues bornées sur $I\!R^d$ on a $[f(X_n)] \to E[f(X)]$.

Comparons à la notion introduite au paragraphe 2-6: si E est fini ou dénombrable, on le munit naturellement de la topologie discrète (chaque singleton est un ouvert), de sorte que toute fonction sur E est continue. Ainsi, au vu de la proposition 2-20, on voit que les définitions 2-19 et 15 ci-dessus coïncident. Il faut toutefois faire attention si E est une partie de $I\!R^d$: les convergences en loi au sens de 2-19 et au sens ci-dessus ne coïncident que si la topologie usuelle sur $I\!R^d$ induit sur E la topologie discrète, ce qui revient à dire que les points de E sont isolés dans $I\!R^d$. Exercice: si ce n'est pas le cas, la convergence au sens ci-dessus entraine-t-elle la convergence au sens de 2-19, ou le contraire ?

Là encore, la convergence en loi est en fait une propriété des lois, et en toute rigueur on devrait écrire que la loi de X_n converge vers celle de X.

Proposition 16: *Soit X_n et X des variables aléatoires réelles de fonctions de répartition respectives F_n et F . Pour que $X_n \stackrel{\mathcal{L}}{\to} X$ il faut et il suffit que $F_n(x) \to F(x)$ pour tout x tel que $F(x-) = F(x)$.*

Noter que l'ensemble $D = \{x : F(x-) = F(x)\}$, i.e. l'ensemble des points où F est continue, est dense dans $I\!R$, et même que son complémentaire est au plus dénombrable.

Preuve. a) Supposons d'abord que $X_n \stackrel{\mathcal{L}}{\to} X$. Soit a avec $F(a-) = F(a)$. Pour tout $p \in I\!N^\star$ et tout $b \in I\!R$, il existe une fonction $f_{p,b} \in C(I\!R)$ telle que

$$1_{]-\infty,b]} \;\leq\; f_{p,b} \;\leq\; 1_{]-\infty,b+1/p]}. \tag{16}$$

On a aussi $E[f_{p,b}(X_n)] \to E[f_{p,b}(X)]$ si $n \uparrow \infty$. De (16) on déduit d'abord que $F_n(a) = P(X_n \leq a) \leq E[f_{p,a}(X_n)]$ et $E[f_{p,a}(X)] \leq F(a + \frac{1}{p})$; donc $\limsup_n F_n(a) \leq F(a + \frac{1}{p})$ pour tout p, donc aussi $\limsup_n F_n(a) \leq F(a)$. On en déduit ensuite que $F_n(a) \geq E[f_{p,a-1/p}(X_n)]$ et $E[f_{p,a-1/p}(X)] \geq F(a - \frac{1}{p})$; donc $\liminf_n F_n(a) \geq F(a - \frac{1}{p})$ pour tout p, donc aussi $\liminf_n F_n(a) \geq F(a)$ puisque $F(a-) = F(a)$. Ces deux résultats impliquent que $F_n(a) \to F(a)$.

b) Inversement, supposons que $F_n(x) \to F(x)$ pour tout $x \in T$, où T est une partie dense de $I\!R$. Soit $f \in C(I\!R)$ et $\varepsilon > 0$. Soit $a, b \in T$ avec $F(a) \leq \varepsilon$ et $F(b) \geq 1 - \varepsilon$. Il existe n_0 tel que

$$n \geq n_0 \quad \Rightarrow \quad P(X_n \notin]a,b]) \;=\; 1 - F_n(b) + F_n(a) \;\leq\; 3\varepsilon. \tag{17}$$

La fonction f est uniformément continue sur $]a, b]$, donc il existe un nombre fini de points $a_0 = a < a_1 < \ldots < a_k = b$ appartenant tous à T et tels que $|f(x) - f(a_i)| \leq \varepsilon$ si

$a_{i-1} \leq x \leq a_i$. Donc

$$g(x) \;=\; \sum_{i=1}^{k} f(a_i) 1_{]a_{i-1},a_i]}(x)$$

vérifie $|f-g| \leq \varepsilon$ sur $]a,b]$. Si $M = \sup_x |f(x)|$, il vient alors

$$|E[f(X_n)] - E[g(X_n)]| \;\leq\; M \;P(X_n \notin\,]a,b]) + \varepsilon, \tag{18}$$

et de même pour X. Enfin $E[g(X_n)] = \sum_{i=1}^{k} f(a_i)(F_n(a_{i-1}) - F_n(a_i))$, et de même pour X, par définition de g. Comme $F_n(a_i) \to F(a_i)$ pour tout i, on en déduit l'existence de n_1 tel que

$$n \geq n_1 \quad \Rightarrow \quad |E[g(X_n)] - E[g(X)]| \;\leq\; \varepsilon. \tag{19}$$

D'après (17), (18) et (19) on a

$$n \geq \sup(n_0, n_1) \quad \Rightarrow \quad |E[f(X_n)] - E[f(X)]| \;\leq\; 3\varepsilon + 4M\varepsilon.$$

ε étant arbitraire, on en déduit que $E[f(X_n)] \to E[f(X)]$, et on a le résultat. $\square$

Théorème 17 (de Lévy): *Soit X_n des variables aléatoires à valeurs dans $I\!R^d$.*

a) Si les $X_n \stackrel{\mathcal{L}}{\to} X$, alors ϕ_{X_n} converge simplement vers ϕ_X.

b) Si les ϕ_{X_n} convergent simplement vers une fonction (complexe) ϕ sur $I\!R^d$, et si cette fonction est **continue** *en 0, alors c'est la fonction caractéritique d'une variable aléatoire X et $X_n \stackrel{\mathcal{L}}{\to} X$.*

Preuve de (a). (Nous ne démontrons pas (b), qui est assez difficile). Il suffit de remarquer que $\phi_{X_n}(u) = E(g_u(X_n))$ et $\phi_X(u) = E(g_u(X))$, où g_u est la fonction continue bornée $g_u(x) = e^{i<u,x>}$ et d'appliquer la définition 15 (en séparant parties réelle et imaginaire). $\square$

Exemples: Ce théorème, plus les formules du paragraphe 1 donnant les fonctions caractéristiques de plusieurs lois, impliquent immédiatement que $X_n \stackrel{\mathcal{L}}{\to} X$ dans les cas suivants:

1) X_n suit $B(p_n, m)$, X suit $B(p, m)$, et $p_n \to p$.

2) X_n suit $\mathcal{N}(m_n, \sigma_n^2)$, X suit $\mathcal{N}(m, \sigma^2)$, et $m_n \to m$, $\sigma_n^2 \to \sigma^2$.

3) X_n et X suivent des lois de Poisson de paramètres θ_n et θ, et $\theta_n \to \theta$.

4) X suit $\Gamma(\alpha_n, \theta_n)$ et X suit $\Gamma(\alpha, \theta)$, et $\theta_n \to \theta$, $\alpha_n \to \alpha$.

5 Convergences de variables aléatoires

La convergence introduite au paragraphe 4 concerne les lois des variables aléatoires considérées: elle signifie que les lois sont asymptotiquement "proches", mais nullement que les variables aléatoires elles-mêmes sont proches. Ci-dessous, nous allons étudier des modes de convergence impliquant la proximité des variables aléatoires elle-mêmes.

On considère une suite $(X_n)_{n\geq 1}$ de variables aléatoires, et une variable aléatoire “limite” X, toutes définies sur le même espace d'états Ω (muni de la tribu $\mathcal{A}$ et de la probabilité P), et toutes à valeurs dans le même espace $I\!R^d$.

Définition 18. a) La suite (X_n) converge **presque sûrement** vers X, ce qui s'écrit $X_n \to X$ p.s., s'il existe un ensemble $N \subset \Omega$ de probabilité nulle, tel que $X_n(\omega) \to X(\omega)$ pour tout $\omega \notin N$.

b) La suite (X_n) converge **en probabilité** vers X, ce qui s'écrit $X_n \stackrel{\mathrm{P}}{\to} X$, si pour tout $\varepsilon > 0$ on a

$$P(|X_n - X| \geq \varepsilon) \;\to\; 0 \qquad \text{quand} \;\; n \to \infty. \tag{20}$$

c) Si $d = 1$, la suite (X_n) converge **en moyenne** vers X, ce qui s'écrit $X_n \stackrel{\mathcal{L}^1}{\to} X$, si X_n et X sont dans $\mathcal{L}^1$ et si

$$E(|X_n - X|) \;\to\; 0 \qquad \text{quand} \;\; n \to \infty. \tag{21}$$

Proposition 19: *La converence p.s. et la convergence en moyenne entraînent la convergence en probabilité.*

Preuve. Soit $A_{n,\varepsilon} = \{|X_n - X| \geq \varepsilon\}$.

a) Supposons que $X_n \to X$ p.s. et soit N l'ensemble de probabilité nulle en dehors duquel on a $X_n(\omega) \to X(\omega)$. Si $\omega \notin N$ on a $\omega \notin A_{n,\varepsilon}$ pour tout $n \geq n_0$, où n_0 dépend de ω et de ε, ce qui implique que les variables aléatoires $Y_{n,\varepsilon} = 1_{N^c \cap A_{n,\varepsilon}}$ tendent simplement vers 0 quand $n \to \infty$. Comme on a aussi $0 \leq f_{n,\varepsilon} \leq 1$, le théorème 2 entraine que $E(Y_{n,\varepsilon}) \to 0$. Mais

$$P(A_{n,\varepsilon}) \;\leq\; P(N^c \cap A_{n,\varepsilon}) + P(N) \;=\; P(N^c \cap A_{n,\varepsilon}) \;=\; E(Y_{n,\varepsilon}) \;\to\; 0,$$

et on en déduit (20).

b) Supposons que $X_n \stackrel{\mathcal{L}^1}{\to} X$. Pour $\varepsilon > 0$ on a $1_{A_{n,\varepsilon}} \leq \frac{1}{\varepsilon}|X - X_n|$, donc

$$P(A_{n,\varepsilon}) \;\leq\; \frac{1}{\varepsilon} E(|X_n - X|) \;\to\; 0,$$

d'où encore (20). $\square$

La convergence en probabilité n'entraîne pas la convergence en moyenne, ne serait-ce que parce qu'elle n'implique pas l'appartenance de X_n et X à $\mathcal{L}^1$. Si les X_n ne sont pas trop grands, il y a cependant équivalence entre les deux modes de convergence. En voici un exemple:

Proposition 20: *S'il existe une constante a telle que $|X_n| \leq a$ identiquement, il y a équivalence entre $X_n \stackrel{P}{\to} X$ et $X_n \stackrel{\mathcal{L}^1}{\to} X$.*

Preuve. Etant donnée la proposition précédente, dont on reprend les notations, il suffit de montrer que la convergence en probabilité implique la convergence en moyenne, lorsque $|X_n| \leq a$.

On a $\{|X| > a+\varepsilon\} \subset A_{n,\varepsilon}$ puisque $|X_n| \leq a$, donc $P(|X| > a+\varepsilon) \leq P(A_{n,\varepsilon})$. En faisant $n \to \infty$ on en déduit que $P(|X| > a+\varepsilon) = 0$. En faisant ensuite tendre ε vers 0, on obtient

$$P(|X| > a) \;=\; 0. \tag{22}$$

Comme $|X_n| \leq a$ on a aussi

$$|X_n - X| \;\leq\; \varepsilon + (|X_n| + |X|)1_{A_{n,\varepsilon}} \;\leq\; \varepsilon + 2a1_{A_{n,\varepsilon}}$$

sur l'ensemble $\{|X| \leq a\}$, qui est de probabilité 1. Donc en utilisant (3-24) il vient

$$E(|X_n - X|) \;\leq\; \varepsilon + 2aP(A_{n,\varepsilon}).$$

On déduit (par (20)) que $\limsup_n E(|X_n - X|) \leq \varepsilon$, et comme ε est arbitrairement proche de 0 on a en fait (21). $\square$

Les rapports entre convergence p.s. et convergence en probabilité sont plus subtils. La première de ces deux convergences est plus forte que la seconde d'après la proposition 19, mais "à peine plus", comme le montre le résultat suivant, donné sans démonstration:

Proposition 21: *Si $X_n \xrightarrow{P} X$, il existe une sous-suite (n_k) telle que $X_{n_k} \to X$ p.s. quand $k \to \infty$.*

Exemple: Soit $\Omega = I\!R$ avec sa tribu borélienne $\mathcal{A} = \mathcal{R}$ et la probabilité P uniforme sur $[0,1]$. Soit $X_n = 1_{A_n}$, où A_n est un intervalle de $[0,1]$ de longueur $1/n$. On a alors $E(X_n) = 1/n$, donc la suite X_n tend vers $X = 0$ en moyenne, et donc en probabilité. Cependant si les A_n sont placés bout-à-bout, en recommençant en 0 chaque fois qu'on arrive au point 1, on voit qu'on parcourt indéfiniment l'intervalle $[0,1]$ (car la série de terme général $1/n$ diverge): donc la suite numérique $X_n(\omega)$ ne converge pas si $\omega \in [0,1/2]$, et on n'a pas $X_n \to X$ p.s.; cependant comme la série $\sum_n 1/n^2$ converge, on voit que $X_{n^2} \to X = 0$ p.s.

Proposition 22: *Soit f une fonction continue de $I\!R^d$ dans $I\!R$.*

a) Si $X_n \to X$ p.s., alors $f(X_n) \to f(X)$ p.s.

b) Si $X_n \xrightarrow{P} X$, alors $f(X_n) \xrightarrow{P} f(X)$.

Preuve. (a) est évident. Pour (b) remarquons d'abord que si $K > 0$ et $\varepsilon > 0$,

$$\{|f(X_n) - f(X)| \geq \varepsilon\} \;\subset\; \{|X| > K\} \cup \{|X| \leq K, |f(X_n) - f(X)| \geq \varepsilon\}. \tag{23}$$

La fonction f est uniformément continue sur $\{x : |x| \leq K\}$, donc il existe $\eta > 0$ tel que $|x-y| < \eta$ et $|x| \leq K$ impliquent $|f(x) - f(y)| < \varepsilon$. Donc (23) implique

$$\{|f(X_n) - f(X)| \geq \varepsilon\} \;\subset\; \{|X| > K\} \cup \{|X_n - X| \geq \eta\},$$

$$P(|f(X_n)-f(X)|\geq\varepsilon) \;\leq\; P(|X|>K)+P(|X_n-X|\geq\eta).$$

D'après l'hypothèse il vient

$$\limsup_n P(|f(X_n)-f(X)|\geq\varepsilon) \;\leq\; P(|X|>K). \tag{24}$$

Enfin $\lim_{K\to\infty} P(|X|>K)=0$, donc dans (24) la lim sup est nulle, et on a le résultat. □

Enfin, toutes les convergences introduites ci-dessus sont plus fortes que la convergence en loi, comme le montre la

Proposition 23: *Si $X_n \stackrel{P}{\to} X$, alors $X_n \stackrel{\mathcal{L}}{\to} X$.*

Preuve. Soit $f\in C(\mathbb{R}^d)$. D'après la proposition 22 on a $f(X_n)\stackrel{\mathrm{P}}{\to} f(X)$, donc $f(X_n)$ converge aussi en moyenne vers $f(X)$ par la proposition 20. Comme $|E(Y)|\leq E(|Y|)$ pour toute variable aléatoire réelle Y, on en déduit qu'*a fortiori* $E[f(X_n)]\to E[f(X)]$. □

6 La loi des grands nombres

Dans ce paragraphe on considère une suite $(X_n)_{n\geq 1}$ de variables aléatoires réelles **indépendantes et de même loi**: dire qu'elles ont indépendantes sous-entend qu'elles sont définies sur le même espace de probabilité. On considère la "moyenne" des n premières variables aléatoires, i.e.

$$M_n \;=\; \frac{1}{n}(X_1+\ldots+X_n), \tag{25}$$

et notre objectif est de montrer que M_n converge vers l'espérance des X_n lorsque cette dernière existe (comme les X_n ont même loi, cette espérance ne dépend pas de n). Il s'agit là d'un des résultats essentiels de toute la théorie des probabilités, connu sous le nom de **loi des grands nombres**.

Commençons par un résultat partiel, mais que nous démontrons complètement.

Théorème 24: *On suppose les X_n dans $\mathcal{L}^2$, et on note $m=E(X_n)$ leur moyenne commune. On a alors*

$$M_n \;\to\; m \qquad \textit{p.s. et en moyenne} \tag{26}$$

(on a donc aussi convergence en probabilité). On a même un peu plus que la convergence en moyenne, à savoir que

$$E[(M_n-m)^2] \;\to\; 0. \tag{27}$$

Preuve. Notons σ^2 la variance de toutes les variables X_n, qui existe puisqu'on a supposé $X_n\in\mathcal{L}^2$. En vertu de la linéarité de l'espérance et de (13) et (3-28), on a

$$E(M_n) \;=\; m, \qquad E[(M_n-m)^2] \;=\; \mathrm{var}(M_n) \;=\; \frac{\sigma^2}{n}, \tag{28}$$

d'où (27). Comme $E(|Y|)^2\leq E(Y^2)$, on en déduit que $E(|M_n-m|)\to 0$, donc on a aussi la convergence de M_n vers m en moyenne. Il reste à montrer la convergence p.s. Quitte à remplacer X_n par X_n-m (donc M_n par M_n-m), on peut supposer que $m=0$.

D'après l'inégalité de Bienaymé-Tchebicheff, (28) implique

$$P(|M_{n^2}| \geq \frac{1}{q}) \;\leq\; \frac{\sigma^2 q^2}{n^2}.$$

Donc si $A_{n,q} = \{|M_{n^2}| \geq \frac{1}{q}\}$, on a $\sum_{n\geq 1} P(A_{n,q}) < \infty$. Posons ensuite $B_{n,q} = \cup_{m\geq n} A_{m,q}$. On a $P(B_{n,q}) \leq \sum_{m=n}^{\infty} P(A_{m,q})$, qui est le reste d'une série convergente. Donc $P(B_{n,q}) \to 0$ quand $n \to \infty$. Si alors $C_q = \cap_{n\geq 1} B_{n,q}$, on a $P(C_q) \leq P(B_{n,q})$ pour tout n, donc en fait $P(C_q) = 0$. Par suite si on pose $N = \cup_{q\in I\!N^*} C_q$, on obtient $P(N) \leq \sum_{q=1}^{\infty} P(C_q) = 0$ en vertu de (1-17).

Maintenant, si $\omega \notin N$, pour tout $q \geq 1$ on a $\omega \notin C_q$, donc aussi $\omega \notin B_{n,q}$ pour n assez grand (car $B_{n,q}$ est décroissant en n). Cela veut dire que pour tout $\omega \notin N$, pour tout $q \geq 1$ il existe n assez grand tel que $M_{m^2}(\omega) \leq \frac{1}{q}$ dès que $m \geq n$. En d'autres termes, $M_{n^2}(\omega) \to 0$ si $\omega \notin N$, donc

$$M_{n^2} \;\to\; 0 \qquad \text{p.s.} \tag{29}$$

Pour tout entier n on note $p(n)$ l'entier tel que $p(n)^2 \leq n < (p(n)+1)^2$. On a

$$M_n - \frac{p(n)^2}{n} M_{p(n)^2} \;=\; \frac{1}{n} \sum_{p=p(n)^2+1}^{n} X_p,$$

et comme pour la seconde égalité (28), on a

$$\begin{aligned} E\left(\left(M_n - \frac{p(n)^2}{n} M_{p(n)^2}\right)^2\right) \;&=\; \frac{n-p(n)^2}{n^2}\sigma^2 \\ &\leq\; \frac{2p(n)+1}{n^2}\sigma^2 \;\leq\; \frac{2\sqrt{n}+1}{n^2}\sigma^2, \end{aligned}$$

parce que $p(n) \leq \sqrt{n}$. D'après Bienaymé-Tchebicheff encore, on a

$$P\left(\left|M_n - \frac{p(n)^2}{n} M_{p(n)^2}\right| \geq a\right) \;\leq\; \frac{2\sqrt{n}+1}{n^2}\;\frac{\sigma^2}{a^2}.$$

Comme la série $\sum_n \frac{2\sqrt{n}+1}{n^2}$ converge, le même raisonnement que ci-dessus pour (29) montre que

$$M_n - \frac{p(n)^2}{n} M_{p(n)^2} \;\to\; 0 \qquad \text{p.s.}$$

Par ailleurs $M_{p(n)^2} \to 0$ p.s. d'après (29), et $p(n)/n^2 \to 1$. On en déduit que $M_n \to 0$ p.s. □

Plus généralement, on a le résultat suivant, admis sans démonstration:

Théorème 25: *On suppose que les X_n sont dans $\mathcal{L}^1$. Avec les mêmes notations que ci-dessus, on a (26).*

Revenons à "l'approche par les fréquences" du chapitre 1. Soit un événement A. On répète l'expérience, et on note X_n la variable aléatoire qui vaut 1 si A est réalisé au cours de la $n^{\text{ème}}$ expérience et 0 sinon. La fréquence de réalisation de A au cours des n premières expériences est alors

$$f_n(A) \;=\; \frac{1}{n}(X_1 + \ldots + X_n) \;=\; M_n.$$

Par ailleurs, les X_i ont même loi et $E(X_i) = P(X_i = 1) = P(A)$, et elles sont indépendantes. Donc (26) implique que $f_n(A) \to P(A)$ p.s.: On obtient ainsi une justification *a posteriori* de l'approche par les fréquences, qui, sans en démontrer de manière rigoureuse la validité (c'est évidemment impossible), montre au moins que cette approche est compatible avec la théorie qui a été basée dessus.

En outre, la loi des grands nombres nous indique aussi dans quel sens il convient de prendre la convergence dans (1-1), à savoir au sens p.s. Il faut remarquer que dans les théorèmes précédents, et donc aussi dans l'approche par les fréquences, on **ne peut pas** avoir convergence de $M_n(\omega)$ vers m pour tout ω: prenons, comme pour l'approche par les fréquences, une suite X_n de variables aléatoires ne prenant que les valeurs 0 et 1. L'espace "minimal" sur lequel on peut définir cette suite est $\Omega = \{0,1\}^{I\!N^*}$: i.e. un point ω est une suite numérique $x_1, \ldots$ de 0 et de 1, et chaque suite est en principe possible. Soit alors P une probabilité sous laquelle les X_n sont indépendantes et de même loi (cf. le théorème 2-18), avec $P(X_n = 1) = p \in]0,1[$. La loi des grands nombres nous dit que pour toute suite $x_1, \ldots$ en dehors d'un ensemble de probabilité nulle, la moyenne $\frac{1}{n}(x_1 + \ldots + x_n)$ tend vers le nombre p. Mais d'une part il existe évidemment beucoup de suites ne vérifiant pas cette propriété (par exemple $x_n = 0$ pour tout n, etc...), et d'autre part chaque suite particulière (y-compris celles qui vérifient cette propriété) est de probabilité nulle.

Ainsi, lorqu'on étudie la convergence des variables aléatoires il est **indispensable** d'introduire la convergence p.s., puisqu'on n'a généralement pas la convergence simple (i.e. pour tout ω).

Une application: Montrons comment on peut appliquer la loi des grands nombres au calcul d'intégrales. Soit à calculer l'intégrale $I = \int_A f(x)dx$, où f est une fonction bornée et A est le cube $\{x = (x_1, \ldots, x_d) : |x_i| \leq \alpha \quad \forall i\}$ de $I\!R^d$. Pour calculer I, on peut simuler une suite $X_1, \ldots, X_n$ de variables aléatoires indépendantes et de loi uniforme sur A: cela revient à dire que si chaque X_n admet les composantes $X_{n,j}$ $(1 \leq j \leq d)$, les variables aléatoires $(X_{n,j} : n \geq 1, 1 \leq j \leq d)$ sont indépendantes et uniformes sur $[-\alpha, \alpha]$. Une valeur approchées de I est alors

$$I_n \;=\; \frac{(2\alpha)^d}{n}\left((f(X_1) + \ldots + f(X_n)\right). \tag{30}$$

En effet la loi uniforme sur A admet la densité $g(x) = \frac{1}{(2\alpha)^d} 1_A(x)$, donc l'espérance des $f(X_i)$ égale $I/(2\alpha)^d$, et il s'ensuit que I_n converge vers I par la loi des grands nombres.

L'inconvénient de cette méthode est que I_n est une approximation "aléatoire" de I, donc on a un peu de peine à contrôler l'erreur $I_n - I$ (l'objet du paragraphe suivant est précisément le contrôle de cette erreur). L'avantage est qu'elle marche même si la fonction f est très irrégulière (alors que les méthodes déterministes de type "méthode du trapèze" ne marchent que si la fonction f est continue). En outre elle marche indépendamment de la dimension d, le temps de calcul étant proportionnel à d (en effet, tirer une variable X de loi uniforme sur A revient à tirer ses d composantes, chacune selon la loi uniforme sur $[0,1]$), alors que les méthodes déterministes ne sont possibles, du point de vue du temps de calcul, que pour d petit, disons $d \leq 3$, puisque le temps de calcul est grosso modo proportionnel à une constante à la puissance d.

7 Le théorème central limite

La situation est la même que dans le paragraphe précédent: les X_n sont des variables aléatoires **indépendantes, de même loi, et dans $\mathcal{L}^2$**. On note m et σ^2 la moyenne et la variance des X_n, et aussi

$$S_n \;=\; X_1 + \ldots + X_n \tag{31}$$

(ainsi $M_n = S_n/n$). On a vu que S_n/n converge vers m p.s. et en moyenne, et il est naturel de chercher la vitesse à laquelle cette convergence a lieu.

Pour évaluer cette vitesse, c'est-à-dire trouver un équivalent de $S_n/n-m$, on est amené à étudier la limite éventuelle de la suite $n^\alpha(S_n/n - m)$ pour différentes valeurs de α: si α est "petit" cette suite va encore tendre vers 0, et elle va "exploser" si α est "grand". On peut espérer que pour une (et alors nécessairement une seule) valeur de α, cette suite converge vers une limite qui n'est ni infinie ni nulle.

Il se trouve que la réponse à cette question a un aspect "négatif": la suite $n^\alpha(S_n/n-m)$ ne converge au sens p.s., ou même en probabilité, pour aucune valeur de α. Elle a aussi un aspect "positif": cette suite converge, au sens de la convergence en loi, pour la même valeur $\alpha = 1/2$ quelle que soit la loi des X_n, et toujours vers une loi normale ! (si $\sigma > 0$, car sinon on a $X_n = m$ et $S_n/n - m = 0$ p.s. pour tout n, et le problème n'a aucun intérêt). Ce résultat, qui peut sembler miraculeux, montre pourquoi la loi normale joue un rôle aussi important en probabilités. Il fait l'objet du théorème suivant, appelé **théorème central limite**, ou **de la limite centrale**:

Théorème 26: *Si les X_n sont des variables aléatoires réelles indépendantes et de même loi, dans $\mathcal{L}^2$, et de moyenne et variance m et σ^2 avec $\sigma^2 > 0$, alors les variables $\frac{S_n-nm}{\sigma\sqrt{n}}$ convergent en loi vers une variable aléatoire de loi $\mathcal{N}(0,1)$.*

En d'autres termes, $\sqrt{n}(S_n/n - m)$ converge en loi vers une variable normale de loi $\mathcal{N}(0,\sigma^2)$.

Preuve. Soit ϕ la fonction caractéristique de $X_n - m$, et $Y_n = (S_n - nm)/\sqrt{n\sigma^2}$. Daprès (5) et (12), la fonction caractéristique de Y_n est

$$\phi_n(u) \;=\; \phi\left(\frac{u}{\sigma\sqrt{n}}\right)^n. \tag{32}$$

Comme $E(X_n - m) = 0$ et $E[(X_n - m)^2] = \sigma^2$, (2) entraine

$$\phi(u) \;=\; 1 - \frac{u^2\sigma^2}{2} + u^2 o(|u|) \qquad \text{quand} \;\; u \to 0.$$

D'après (32) on a alors pour chaque u fixé et tout n assez grand pour que $|\phi\left(\frac{u}{\sigma\sqrt{n}}\right) - 1| \leq 1/2$:

$$\phi_n(u) \;=\; \exp n \log\left(1 - \frac{u^2}{2n} + \frac{1}{n}\varepsilon_n(u)\right),$$

où $\varepsilon_n(u) \to 0$ quand $n \to \infty$, et où $\log z$ désigne la détermination principale du logarithme du nombre complexe z, qui est bien définie sur le disque $\{z \in \mathbb{C} : |z-1| \leq 1/2\}$ et qui sur ce disque admet le même développement limité au voisinage de $z = 1$ que le logarithme réel. On en déduit alors immédiatement que $\phi_n(u) \to \exp -u^2/2$, et le résultat découle du théorème 17. $\square$

Remarque: Il est facile de déduire de ce résultat que $n^\alpha |S_n/n - m|$ converge vers 0 (resp. $+\infty$) en probabilité lorsque $\alpha < 1/2$ (resp. $\alpha > 1/2$).

Exemple: convergence des lois binomiales. Supposons que S_n suive une loi binomiale $B(p,n)$. Cela revient à dire que S_n a la même loi qu'une somme $X_1 + \ldots + X_n$ de n variables aléatoires X_i indépendantes de loi $B(p,1)$, i.e. $P(X_i = 1) = p$ et $P(X_i = 0) = 1-p$. On a alors $m = p$ et $\sigma^2 = p(1-p)$. Donc en vertu des théorèmes précédents,

$$\frac{S_n}{n} \overset{\mathrm{P}}{\to} p, \tag{33}$$

$$\frac{S_n - np}{\sqrt{np(1-p)}} \overset{\mathcal{L}}{\to} \mathcal{N}(0,1). \tag{34}$$

Supposons alors qu'on doive calculer $P(S_n \leq x)$ pour n grand. Si p est très petit, de sorte que $\theta = np$ ne soit pas trop grand (en pratique, $\theta \leq 5$ convient), on peut utiliser l'approximation par une loi de Poisson du paragraphe 2-2. Si p est très proche de 1, de sorte que $\theta = n(1-p)$ soit comme ci-dessus, alors $n - S_n$ suit à son tour une loi proche de la loi de Poisson de paramètre θ. Dans les autres cas, on utilise (34): F désignant la fonction de répartition $\mathcal{N}(0,1)$, on a

$$P(S_n \leq x) \;\sim\; F\left(\frac{x - np}{\sqrt{np(1-p)}}\right). \tag{35}$$

Vu l'importance de la loi normale, la fonction de répartition F ci-dessus a été tabulée, et une table de cette fonction est fournie en annexe de ce cours.

Le thórème 26 admet une version multidimensionnelle, dont la preuve est rigoureusement identique. On suppose que les X_n sont des variables aléatoires à valeurs dans $\mathbb{R}^d$, indépendantes et de même loi. On suppose que les composantes des X_n sont dans $\mathcal{L}^2$. On a ainsi un vecteur moyenne $m = E(X_n)$, et une matrice de covariance $C = (c_{ij})$ avec c_{ij} = la covariance des composantes i et j de X_n. On a alors le:

Théorème 27: *Les variables aléatoires $\frac{S_n - nm}{\sqrt{n}}$ convergent en loi vers un vecteur aléatoire gaussien centré (i.e. de moyenne nulle), de matrice de covariance C.*

CHAPITRE 5

Statistiques: l'estimation

1 Introduction

Le problème de la statistique est le suivant; on observe un certain nombre de variables aléatoires, dont les lois sont - complètement ou partiellement - inconnues. Il s'agit, à partir des observations, d'obtenir le plus possible d'informations sur ces lois.

Par exemple, supposons qu'on fabrique des pièces sur une machine. Chaque pièce fabriquée a une probabilité θ inconnue, mais la même pour toutes les pièces, d'être défectueuse: ce nombre θ dépend du réglage de la machine, le réglage est d'autant meilleur que θ est proche de 0; mais comme le réglage ne peut être parfait on n'a jamais $\theta = 0$. Avant de lancer le cycle de fabrication, on veut vérifier si la machine est "bien réglée", i.e. si θ est suffisamment petit (rappelons qu'il ne peut être exactement nul !). Pour cela on fabrique un certain nombre n de pièces qui servent à tester le réglage. L'observation consiste à compter le nombre X de pièces défectueuses parmi ces n pièces. On peut alors se poser deux types de problèmes:

1) Trouver "la" valeur de θ: cela s'appelle **estimer** le paramètre θ. Dans notre exemple, il est naturel de prendre pour **estimateur** de θ la proportion $\widehat{\theta}_n = X/n$ de pièces défectueuses.

2) S'assurer que la vraie valeur de θ ne dépasse un seuil critique θ_0 fixé à l'avance (sinon, il faut refaire le réglage de la machine): cela s'appelle **tester** le fait que $\theta \leq \theta_0$.

Ces deux problèmes sont de nature mathématique assez différente. Ils ont cependant en commun le fait **qu'on ne peut pas arriver à une conclusion certaine**: dans le cas (1) il est "vraisemblable" que la valeur exacte de θ soit proche de l'estimation $\widehat{\theta}_n$ (au moins si n est assez grand), mais tout-à-fait invraisemblable qu'elle lui soit exactement égale. Dans le cas (2) on peut décider que $\theta \leq \theta_0$ si la proportion X/n est suffisamment petite, mais on ne sera jamais sûr que la vraie valeur de θ soit effectivement plus petite que le seuil θ_0.

On peut toujours représenter un problème de statistique de la manière suivante:

Définition 1. On appelle **modèle statistique** la donnée de:

- un espace d'états Ω, qui est l'ensemble de tous les résultats possibles de l'expérience réalisée; comme dans les chapitres précédents, cet espace est muni de la tribu $\mathcal{A}$ des événements;

- une famille $(P_\theta)_{\theta\in\Theta}$ de probabilités sur $(\Omega, \mathcal{A})$.

Enfin, on appelle **statistique** toute variable aléatoire sur cet espace.

Dans tous les cas, on cherche, à partir de la connaissance de ω (aléatoire, représentant l'observation), à obtenir des renseignements sur la valeur inconnue (mais non aléatoire) du paramètre θ:

1) Soit on veut déterminer la valeur de θ, ou d'une fonction $f(\theta)$; il s'agit alors **d'estimation**.

2) Soit on veut savoir si θ se trouve dans une partie Θ_0 de l'ensemble Θ, ou dans son complémentaire: il s'agit alors d'un **test**.

Exemple A: Reprenons l'exemple introductif de la fabrication de pièces. On observe le nombre X de pièces défectueuses, donc $\Omega = \{0, 1, \dots, n\}$, avec la tribu $\mathcal{A} = \mathcal{P}(\Omega)$, et $X(\omega) = \omega$. L'ensemble des paramètres est $\Theta =]0, 1[$, et pour $\theta \in \Theta$, la probabilité P_θ est la loi binomiale $B(\theta, n)$.

Exemple B: On veut mesurer une longueur inconnue u, et à cet effet on prend n mesures successives, dont les résultats sont $X_1, \dots, X_n$. Le modèle est constitué de $\Omega =]0, \infty[^n$ muni de la tribu borélienne $\mathcal{A}$, et des variables aléatoires $X_i(x_1, \dots, x_n) = x_i$ si $\omega = (x_1, \dots, x_n) \in \Omega$. Quant aux probabilités, il est naturel de supposer que les X_i sont indépendantes, de même loi μ admettant pour moyenne m la quantité u à mesurer: en termes de modèle statistique, on est donc conduit à considérer pour Θ l'ensemble de toutes les probabilités sur $]0, \infty[$, et P_θ sera l'unique probabilité sur Ω pour laquelle les X_i sont indépendantes et de loi θ. L'espace Θ est donc très gros, mais on ne s'intéresse qu'à la fonction $f(\theta) = \int x\theta(dx)$, qui est la moyenne de la loi θ et qui est supposée être égale à la quantité u qu'on cherche à mesurer.

On voit sur ces deux exemples les deux types extrêmes de modèles statistiques: dans un cas l'espace Θ est une partie de $\mathbb{R}$, ou plus généralement de $\mathbb{R}^d$: on dit qu'on a un problème **paramétrique**. Dans l'autre cas, Θ est l'espace de **toutes** les probabilités sur un ensemble donné: on dit qu'on a un problème **non-paramétrique**. Dans ce cours nous considérerons essentiellement des problèmes paramétriques.

Enfin, très souvent les problèmes statistiques se posent dans le cadre suivant:

Définition 2. Un n-**échantillon** est une suite $(X_1, \dots, X_n)$ de n variables aléatoires indépendantes et de même loi, à valeurs dans l'espace E (en général $E = \mathbb{Z}$, $E = \mathbb{R}$ ou $E = \mathbb{R}^d$). On lui associe le modèle suivant: on a une famille $(\mu_\theta)_{\theta\in\Theta}$ de probabilités sur E, et on pose $\Omega = E^n$, qu'on munit des probabilités P_θ sous lesquelles les variables aléatoires $X_i(x_1, \dots, x_n) = x_i$ forment un n-échantillon de loi μ_θ.

Exemples:

- L'exemple B ci-dessus est un exemple de modèle basé sur un n-échantillon.

- L'exemple A est un exemple avec un 1-échantillon. On peut cependant le voir comme un modèle à n-échantillon, en notant X_i la variable aléatoire qui vaut 1 si la $i^{\text{ème}}$ pièce est défectueuse et $X_i = 0$ sinon. Dans cette version du modèle on a $\Omega = \{0,1\}^n$ et aussi, pour $\omega = (i_1, \ldots, i_n) \in \Omega$ et $X(\omega) = X_1(\omega) + \ldots + X_n(\omega) = i_1 + \ldots + i_n$:

$$P_\theta(\{\omega\}) \;=\; \theta^{X(\omega)}(1-\theta)^{n-X(\omega)}.$$

2 Estimateurs: définition, risque quadratique

Dans ce paragraphe on suppose donné le modèle statistique $(\Omega, \mathcal{A}, (P_\theta)_{\theta\in\Theta})$. Soit f une fonction (connue) sur Θ, que pour simplifier on suppose à valeurs réelles. On cherche à **estimer** la quantité (inconnue) $f(\theta)$.

Estimer $f(\theta)$ veut dire qu'au vu de l'observation ω on "décide" que la valeur de $f(\theta)$ vaut un certain nombre, disons $T(\omega)$, qui dépend en général de ω. Autrement dit, on choisit une variable aléatoire réelle, ou une "statistique", T; dans le cadre de l'estimation on appelle aussi T un **estimateur** de $f(\theta)$.

Exemples.

A (suite): Dans la première modélisation où $\Omega = \{0, \ldots, n\}$ et $P_\theta = B(\theta, n)$, la variable aléatoire $T(\omega) = \omega/n$ est un estimateur de θ (on verra que c'est "le meilleur" au sens des critères que nous définirons plus bas). Le carré T^2 est un estimateur raisonnable de θ^2, mais ce n'est pas le meilleur au sens de ces mêmes critères. On peut aussi considérer T^2 (ou T^3, etc...) comme un estimateur de θ, mais il sera bien-sûr "mauvais".

C : Supposons qu'on observe un n-échantillon de la loi normale $\mathcal{N}(\theta, 1)$. Un estimateur raisonnable de θ est la **moyenne empirique**

$$\bar{X} \;=\; \frac{1}{n}(X_1 + \ldots + X_n). \tag{1}$$

Le problème de l'estimation consiste à optimiser le choix de l'estimateur, et pour cela il faut introduire un critère de qualité. *A priori* on a envie de dire que l'estimateur S est meilleur que T si l'erreur commise par S est plus petite en valeur absolue que celle commise par T: il convient alors de remarquer que l'erreur est $T(\omega) - f(\theta)$; elle dépend à la fois du paramètre **inconnu** θ, et du résultat ω de l'expérience, connu mais **aléatoire**; ainsi, deux estimateurs S et T ne seront pratiquement jamais comparables, au sens où par exemple $|S(\omega) - f(\theta)| \leq |T(\omega) - f(\theta)|$ pour tous ω et θ.

L'idée sous-jacente aux divers critères de qualité possibles consiste alors à choisir ce qu'on appelle parfois une "fonction de perte", c'est-à-dire une fonction h de $\mathbb{R}_+$ dans lui-même, qui est croissante et nulle en 0; la "perte" de l'estimateur T est alors $H_{T,\theta}(\omega) = h(|T(\omega) - f(\theta)|)$, et le "risque" associé est l'espérance de $H_{T,\theta}$ par rapport à P_θ, i.e. $R_T(\theta) = E_\theta(H_{T,\theta})$: c'est une fonction de θ, mais elle ne dépend plus de l'aléa ω, et un estimateur S est dit meilleur qu'un autre estimateur T si leurs fonctions de risque satisfont $R_S(\theta) \leq R_T(\theta)$ pour tout θ.

Il faut bien voir le caractère arbitraire de ces critères: d'une part, pour h fixé, dire que S est meilleur que T revient à dire qu'en "moyenne", si on répète souvent l'expérience statistique, la fonction de perte de S sera effectivement plus petite que celle de T, mais ce n'est évidemment pas nécessairement le cas pour **une** expérience donnée. Par ailleurs le choix de h est également arbitraire: si on prend par exemple une fonction puissance, $h(x) = x^\alpha$, plus α est grand plus on privilégie les "grandes" erreurs par rapport aux "petites". Pour des raisons de commodité mathématique, on utilise en général la fonction $h(x) = x^2$, ce qui conduit à la définition suivante:

Définition 3. Le **risque quadratique** de l'estimateur T de $f(\theta)$ est

$$R_T(\theta) \;=\; E_\theta[(T - f(\theta))^2], \tag{2}$$

où E_θ désigne l'espérance au sens de la probabilité P_θ (rappelons que l'espérance dépend de la probabilité, puisque par exemple on a $E(1_A) = P(A)$ pour tout $A \in \mathcal{A}$).

Si S et T sont deux estimateurs de $f(\theta)$, on dit que S est **meilleur** que T (au sens du risque quadratique) si $R_S(\theta) \leq R_T(\theta)$ pour tout $\theta \in \Theta$; il est dit **strictement meilleur** s'il est meilleur, et si de plus $R_S(\theta) < R_T(\theta)$ pour au moins une valeur de θ.

Exemple C (suite): Comme $f(\theta) = \theta$ est la moyenne $E_\theta(\bar{X})$, le risque $R_{\bar{X}}(\theta)$ est la variance de $\bar{X}$ sous P_θ. Comme $\bar{X}$ suit la loi $\mathcal{N}(\theta, 1/n)$, on a $R_{\bar{X}}(\theta) = 1/n$.

X_1 peut aussi être considéré comme un estimateur de θ, de moyenne θ et de risque quadratique $R_{X_1}(\theta) = 1$. Donc $\bar{X}$ est strictement meilleur que X_1 dès que $n \geq 2$.

Remarque 4: La relation "S est meilleur que T " est une relation d'ordre **partiel** sur la famille de tous les estimateurs (i.e. sur la famille de toutes les variables aléatoires). Deux estimateurs donnés ne sont en général pas comparables, et il n'existe pas d'estimateur meilleur que tous les autres. Dans l'exemple C, soit $T(\omega) = a$ pour tout ω, où a et une constante arbitraire; le risque quadratique pour estimer $f(\theta)$ est alors évidemment $R_T(\theta) = (a - \theta)^2$, donc $R_T(\theta) < R_{\bar{X}}(\theta)$ pour certaines valeurs de θ, et on a l'inégalité inverse pour d'autres valeurs de θ. Dans ce cas, T n'est pas un estimateur raisonnable de θ, puisqu'il ne dépend pas de l'observation, cependant son risque est nul quand le paramètre inconnu θ vaut a.

La détermination d'**un meilleur estimateur**, i.e. tel qu'il n'en existe pas de strictement meilleur, est un problème mathématique extrêmement difficile, car la classe de tous les estimateurs est trop vaste. Dans la plupart des cas on se restreint à une classe particulière d'estimateurs, et notamment à la classe suivante:

Définition 5. a) L'estimateur T de $f(\theta)$ est dit **sans biais** (ou, non-biaisé) si

$$E_\theta(T) \;=\; f(\theta), \qquad \forall\theta \in \Theta \tag{3}$$

(dans ce cas le risque quadratique $R_T(\theta)$ est la variance de T sous P_θ).

b) On dit que T est **un meilleur estimateur sans biais** de $f(\theta)$ s'il est sans biais et s'il est meilleur que tout autre estimateur sans biais.

Justification: Il n'existe pas de justification pleinement satisfaisante à l'usage d'estimateurs sans biais. Toutefois, considérons le cas où on observe un n-échantillon $(X_1, \ldots, X_n)$ de loi μ_θ. On veut estimer $f(\theta)$, et on suppose qu'il existe une fonction g telle que

$$E_\theta[g(X_i)] \;=\; f(\theta), \qquad \forall\theta \in \Theta. \tag{4}$$

Alors $T_n = \frac{1}{n}(g(X_1) + \ldots + g(X_n))$ est un estimateur sans biais de $f(\theta)$. Et, d'après la loi des grands nombres, $T_n \to f(\theta)$ P_θ-p.s. lorsque $n \to \infty$. Plus généralement, on peut montrer que sous des conditions assez faibles, les estimateurs sans biais construits sur un n-échantillon convergent vers la valeur à estimer lorsque $n \to \infty$ (mais bien-sûr, dans la pratique, n est fini !).

Exemples:

A (suite): T est un estimateur sans biais de θ.

C (suite): $\bar{X}$ est un estimateur sans biais de θ; par contre $T = a$ est un estimateur biaisé.

D : Supposons qu'on observe une variable aléatoire de loi exponentielle de paramètre $\theta > 0$. On a alors $\Omega = I\!R_+$, et $\Theta =]0, \infty[$, et P_θ est la loi exponentielle de paramètre θ, et on observe la variable aléatoire $X(\omega) = \omega$. Si T était un estimateur sans biais de θ, on aurait

$$\int_0^\infty e^{-\theta x} T(x)dx \;=\; 1 \qquad \forall\theta > 0,$$

ce qui n'est possible pour aucune fonction T: dans ce cas, il n'existe aucun estimateur sans biais.

E **Meilleur estimateur linéaire sans biais**. Considérons le cas où on observe un n-échantillon $X_1, \ldots, X_n$ de loi μ_θ sur $I\!R$ (donc $\Omega = I\!R^n$ et $X_i(\omega) = x_i$ si $\omega = (x_1, \ldots, x_n)$ et P_θ est l'unique probabilité sous laquelle les X_i sont indépendantes, de loi μ_θ). Notons m_θ et σ_θ^2 la moyenne et la variance (supposées exister) de la loi μ_θ. On dit qu'un estimateur est **linéaire** s'il est de la forme

$$T \;=\; b + \sum_{i=1}^n a_i X_i, \tag{5}$$

pour des constantes b et a_i. On cherche à estimer $f(\theta) = m_\theta$. On suppose que m_θ prend au moins deux valeurs (sinon, il n'y a pas de problème !). On a $E_\theta(T) = b + \sum_{i=1}^n a_i f(\theta)$, donc T est sans biais si et seulement si

$$b = 0, \qquad \sum_{i=1}^n a_i = 1.$$

Dans ce cas, le risque quadratique est

$$R_T(\theta) \;=\; \left(\sum_{i=1}^n a_i^2\right) \sigma_\theta^2.$$

Comme $\sum_{i=1}^n a_i = 1$, on a toujours $\sum_{i=1}^n a_i^2 \geq 1/n$, avec égalité si et seulement si $a_i = 1/n$ pour tout i, auquel cas $T = \bar{X}$ est la moyenne empirique (cf. (1)). On a donc démontré la

Proposition 6: *Pour un n-échantillon admettant une variance, et pour estimer la moyenne, la moyenne empirique $\bar{X}$ est le meilleur estimateur parmi tous les estimateurs linéaires sans biais.*

3 Statistiques suffisantes

Dans ce paragraphe on se donne encore un modèle statistique $(\Omega, \mathcal{A}, (P_\theta)_{\theta\in\Theta})$. On se place en outre dans l'un des deux cas suivants:

Cas (a): Ω est fini ou dénombrable.

Cas (b): $\Omega = \mathbb{R}^d$.

Définition 7. Une statistique T à valeurs dans $\mathbb{R}^p$ est dite **suffisante**, ou **exhaustive**, s'il existe une fonction h de Ω dans $\mathbb{R}_0$ et des fonctions q_θ de $\mathbb{R}^q$ dans $]0, \infty[$ telles que

- dans le cas (a), on ait
$$P_\theta(\{\omega\}) \;=\; q_\theta(T(\omega))h(\omega), \tag{6}$$
- dans le cas (b), pour chaque θ la probabilité P_θ admette la densité suivante sur $\mathbb{R}^d$:
$$q_\theta(T(\omega))h(\omega). \tag{7}$$

Cette définition présente un intérêt à cause du résultat suivant, que nous admettrons sans démonstration.

Proposition 8: *Soit T une statistique suffisante. Soit S un estimateur (resp. un estimateur sans biais) de $f(\theta)$. Il existe alors un estimateur de la forme $S' = g(T)$ qui est meilleur que S (resp. et en plus sans biais).*

Ce résultat montre pour estimer $f(\theta)$ on ne peut de toutes façons pas faire mieux que prendre un estimateur qui est une fonction de T: cela éclaire bien la signification des termes “suffisant” et “exhaustif”.

Exemples:

A (suite): Dans la “seconde version” de l'exemple A, on observe un n-échantillon de la loi binomiale $B(\theta, 1)$. On a $\Omega = \{0,1\}^n$, et
$$P_\theta(\{(x_1, \ldots, x_n)\}) \;=\; \theta^{\Sigma\ x_j}\ (1-\theta)^{n-\Sigma\ x_j}.$$
Par suite on a (6) avec $T = \bar{X} = X_1 + \ldots + X_n/n$ et $h(\omega) = 1$ et $q_\theta(x) = \theta^{nx}(1-\theta)^{n(1-x)}$. Donc $\bar{X}$ est suffisante. Du point de vue de l'estimation, on peut modéliser avec un n-échantillon, ou aussi bien avec simplement l'observation de $X_1 + \ldots + X_n$, donc un 1-échantillon de la loi binomiale $B(\theta, n)$ (ce qui correspond à la “première version” de l'exemple A).

F n-**échantillon de loi géométrique:** Le paramètre θ de la loi géométrique appartient à $]0,1[$. On a $\Omega = \mathbb{N}^n$, et

$$P_\theta(\{(x_1,\ldots,x_n)\}) \;=\; (1-\theta)^n \theta^{\sum x_j}.$$

On a donc (6) avec $T = \bar{X} = (X_1 + \ldots + X_n)/n$ et $h(\omega) = 1$ et $q_\theta(x) = (1-\theta)^n \theta^{nx}$. Donc la moyenne empirique $\bar{X}$ est suffisante.

C (suite): On a un n-échantillon de loi normale $\mathcal{N}(\theta, \sigma^2)$, avec σ^2 connu, et θ est le paramètre inconnu. Alors $\Omega = \mathbb{R}^n$ et P_θ admet la densité suivante sur $\mathbb{R}^n$:

$$\frac{1}{(\sqrt{2\pi\sigma^2})^n}\; \exp -\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \theta)^2. \tag{8}$$

On a donc (6) avec $T = \bar{X}$ et $q_\theta(t) = e^{(2n\theta t - n\theta^2)/2\sigma^2}$ et

$$h(x_1,\ldots,x_n) \;=\; \frac{1}{(\sqrt{2\pi\sigma^2})^n}\; \exp -\frac{1}{2\sigma^2}\sum_{i=1}^n x_i^2.$$

Donc $\bar{X}$ est encore une statistique suffisante.

G n-**échantillon de loi normale $\mathcal{N}(m, \sigma^2)$, avec m et σ^2 inconnus:** On a ici $\Theta = \mathbb{R} \times]0,\infty[$ et $\theta = (m, \sigma^2)$, et $\Omega = \mathbb{R}^n$. La densité de la loi $P_\theta = P_{(m,\sigma^2)}$ sur $\mathbb{R}^n$ est toujours donnée par (8). On a donc (6) à condition de prendre pour T la statistique 2-dimensionnelle $T = (\bar{X}, \Sigma^2)$, où $\bar{X}$ est encore la moyenne empirique, et

$$\Sigma^2 \;=\; \frac{1}{n}(X_1^2 + \ldots + X_n^2), \tag{9}$$

et $h = 1$ et

$$q_\theta(u,v) \;=\; \frac{1}{(\sqrt{2\pi\sigma^2})^n}\; \exp -\frac{1}{2\sigma^2}(v - 2nmu + nm^2).$$

Ainsi, le couple $T = (\bar{X}, \Sigma^2)$ est une statistique suffisante.

H n-**échantillon de loi exponentielle de paramètre $\theta \in]0,\infty[$:** On a $\Omega = \mathbb{R}^n$ et P_θ admet la densité suivante sur $\mathbb{R}^n$:

$$\theta^n \; e^{-\theta \sum_i x_i} \; 1_{\mathbb{R}_+^n}(x_1,\ldots,x_n),$$

qui peut clairement se mettre sous la forme (6) avec la statistique suffisante $T = \bar{X}$ et $h(x) = 1_{\mathbb{R}_+^n}(x_1,\ldots,x_n)$ et $q_\theta(x) = \theta^n e^{-n\theta x}$.

4 Les modèles exponentiels

Définition 9. Un **modèle exponentiel** est un modèle avec Ω fini ou dénombrable (cas (a)), ou $\Omega = \mathbb{R}^d$ (cas (b)), admettant une statistique suffisante T à valeurs dans $\mathbb{R}^m$, et tel que les fonctions q_θ de (6) ou (7) se mettent sous la forme

$$q_\theta(t) \;=\; e^{<\alpha(\theta),t> + \beta(\theta)}, \tag{10}$$

où $< .,. >$ désigne le produit scalaire dans $I\!R^m$, où α est une fonction de Θ dans $I\!R^m$, et où β est une fonction de Θ dans $I\!R$.

L'intérêt de ces modèles est dans le théorème suivant, admis sans démonstration.

Théorème 10: *Supposons le modèle exponentiel, et supposons aussi que l'image $\alpha(\Theta)$ de Θ par α contienne un ouvert de $I\!R^m$. Alors, pour n'importe quelle fonction réelle sur $I\!R^m$, la statistique $g(T)$ est le meilleur estimateur sans biais de la fonction $f(\theta) = E_\theta[g(T)]$, dès que $g(T)$ est dans $\mathcal{L}^1$ pour toutes les probabilités P_θ.*

Ce résultat peut sembler tout-à-fait abstrait, mais nous allons en voir une série d'exemples montrant qu'en fait il permet de résoudre un assez grand nombre de problèmes concrets d'estimation. Auparavant, voici une remarque importante pour la pratique, car elle permet souvent de ramener l'étude des n-échantillons à celle des 1-échantillons, *a priori* plus simple.

Remarque 11: Soit $(\mu_\theta)_{\theta\in\Theta}$ une famille de probabilités sur un espace E, ayant une structure exponentielle: cela signifie qu'il existe une fonction h de E dans $I\!R_+$, une application S de E dans $I\!R^m$, une application α_1 de Θ dans $I\!R^m$, et une fonction β_1 de Θ dans $I\!R$, telles que:

- Dans le cas (a) où E est fini ou dénombrable, on a
$$\mu_\theta(\{x\}) \;=\; h(x)e^{<\alpha_1(\theta),S(x)>+\beta_1(\theta)}. \tag{11}$$
- Dans le cas (b) où $E = I\!R^d$, μ_θ admet une densité de la forme
$$h(x)e^{<\alpha_1(\theta),S(x)>+\beta_1(\theta)}. \tag{12}$$

(Cela veut dire simplement que le modèle statistique $(E,(\mu_\theta))$ est un modèle exponentiel).

Considérons alors un n-échantillon $(X_1,\ldots,X_n)$ de loi μ_θ, et le modèle statistique associé. Il est alors très facile de voir que ce modèle aussi est **exponentiel**, avec la statistique T_n et les fonctions α_n et β_n données par
$$\left.\begin{array}{rcl} T_n(\omega) &=& \frac{1}{n}(S(X_1(\omega)) + \ldots + S(X_n(\omega))), \\ \alpha_n(\theta) &=& n\alpha_1(\theta), \qquad \beta_n(\theta) \;=\; n\beta_1(\theta). \end{array}\right\} \tag{13}$$
Par suite si pour un nombre $b \in I\!R$ et un vecteur $a \in I\!R^m$ donnés la statistique $< a,S > +b$ est un estimateur sans biais de $f(\theta) = \int[< a,S(x) > +b]\mu_\theta(dx)$ (donc, d'après le théorème précédent, est aussi **le meilleur** estimateur sans biais de $f(\theta)$ pour un 1-échantillon), on voit que pour le n-échantillon, la statistique $< a,T_n > +b$ est le meilleur estimateur sans biais de $f(\theta)$. $\square$

Voici maintenant des exemples.

Exemples:

A (suite): **Loi binomiale** $\mu_\theta = B(\theta,N)$ avec $\Theta =]0,1[$. On a
$$\mu_\theta(\{i\}) \;=\; C_N^i \;\exp\left(i\log\frac{\theta}{1-\theta} + N\log(1-\theta)\right).$$

On a donc (11) avec $S(i) = i$. Donc S/N est le meilleur estimateur sans biais de θ (et aussi une statistique suffisante, ce qu'on savait déjà). Si maintenant on a un n-échantillon $X_1, \ldots, X_n$ de cette loi, $T = \frac{1}{Nn}(X_1 + \ldots + X_n)$ est le meilleur estimateur sans biais de θ; noter que $X_1 + \ldots + X_n$ suit la loi $B(\theta, nN)$, de sorte que le risque quadratique de T est $R_T(\theta) = \frac{\theta(1-\theta)}{nN}$, qui tend vers 0 lorsque $nN \to \infty$.

I **Loi de Poisson:** Soit μ_θ la loi de Poisson de paramètre $\theta > 0$, sur $E = I\!N$. On a (11) avec $h(n) = 1/n!$, $S(n) = n$, $\alpha_1(\theta) = \log\theta$, $\beta_1(\theta) = -\theta$, et $\sum_n S(n)\mu_\theta(\{n\}) = \theta$. Donc pour un n-échantillon la moyenne empirique $\bar{X}$ est le meilleur estimateur sans biais de θ.

J **Loi gamma:** Soit $\mu_\theta = \Gamma(\alpha, \theta)$, d'indice $\alpha > 0$ fixé et de paramètre $\theta > 0$ inconnu. On a alors (12) avec $S(x) = x$ (prendre aussi $h(x) = \frac{1}{\Gamma(\alpha)} x^{\alpha-1} 1_{]0,\infty[}(x)$, $\alpha_1(\theta) = -\theta$ et $\beta_1(\theta) = \log\theta$). On a $\int S(x)\mu_\theta(dx) = \alpha/\theta$. On en déduit que pour un n-échantillon, la moyenne empirique est le meilleur estimateur sans biais de α/θ.

C (suite): **Loi normale de variance fixée.** On a $\mu_\theta = \mathcal{N}(\theta, \sigma^2)$ avec σ^2 connu et $\Theta = I\!R$. On a (12) avec $S(x) = x$, $\alpha_1(\theta) = \theta/\sigma^2$, $\beta_1(\theta) = -\theta^2/2\sigma^2$ et $h(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp -x^2/2\sigma^2$. On a aussi $\int S(x)\mu_\theta(dx) = \theta$, donc pour un n-échantillon la moyenne empirique est le meilleur estimateur sans biais de la moyenne θ.

K **Loi normale de moyenne fixée.** On a $\mu_\theta = \mathcal{N}(m, \theta)$ avec m connu et $\Theta =]0, \infty[$. On a (12) avec $S(x) = (x-m)^2$ (c'est bien une statistique, car m est connu), $\alpha_1(\theta) = -1/2\theta$, $\beta_1(\theta) = -\log\sqrt{2\pi\theta}$ et $h = 1$. On a aussi $\int S(x)\mu_\theta(dx) = \theta$, donc pour un n-échantillon la statistique

$$T_n = \frac{1}{n}\Big((X_1 - m)^2 + \ldots + (X_n - m)^2\Big) \tag{14}$$

est le meilleur estimateur sans biais de la variance $\theta = \sigma^2$.

G (suite): **Loi normale de moyenne et variance inconnues.** Soit $\mu_\theta = \mathcal{N}(m, \sigma^2)$ avec $\theta = (m, \sigma^2)$ et $\Theta = I\!R \times]0, \infty[$. On a (12) avec $S = (S_1, S_2)$, où

$$S_1(x) = x, \qquad S_2(x) = x^2,$$

et $\alpha_1(\theta) = (m/\sigma^2, -1/2\sigma^2)$, $\beta_1(\theta) = -\log\sqrt{2\pi\sigma^2} - m^2/2\sigma^2$ et $h = 1$. On en déduit que pour un n-échantillon le couple $(\bar{X}, \Sigma^2)$ est suffisant (pour Σ^2, voir (9)); on avait déjà vu cette propriété), et comme $\int S_1(x)\mu_\theta(dx) = m$, la moyenne empirique est encore le meilleur estimateur sans biais de m.

Cherchons maintenant le meilleur estimateur sans biais de σ^2. D'après le théorème 10, c'est l'unique fonction $U = f(\bar{X}, \Sigma^2)$ qui est un estimateur sans biais de σ^2. Remarquons que (14) ne définit pas ici un estimateur, car m est inconnu. Un candidat naturel est alors

$$V = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \Sigma^2 - \bar{X}^2. \tag{15}$$

Cependant, on a

$$E_\theta(V) = E_\theta[(X_1 - \bar{X})^2] = E_\theta\left[\left(\frac{n-1}{n}X_1 - \frac{1}{n}X_2 - \ldots - \frac{1}{n}X_2\right)^2\right] = \frac{n-1}{n}\sigma^2,$$

et V est biaisé. Mais on déduit du calcul précédente que $\frac{n}{n-1}V$ est un estimateur sans biais de σ^2. Donc la statistique suivante, appelée **variance empirique**:

$$S^2 \;=\; \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2 \;=\; \frac{n}{n-1}V \tag{16}$$

est le meilleur estimateur sans biais de σ^2.

5 Intervalles de confiance

Comme nous l'avons déjà souligné, l'erreur $T(\omega)-f(\theta)$ commise en remplaçant $f(\theta)$ par $T(\omega)$ est à la fois aléatoire et dépendante du paramètre inconnu θ. Le risque quadratique est une mesure "deterministe" de cette erreur (ou plutôt, de son carré), mais il dépend encore de la valeur inconnue θ. On préfère donc souvent utiliser la notion suivante, plus parlante, et qui donne une "fourchette d'estimation":

Définition 12. Soit T un estimateur de $f(\theta)$, et $\alpha\in]0,1[$ (on fixe *a priori* ce nombre, proche de 1: typiquement $0,9$ ou $0,95$ ou $0,99$). On appelle **intervalle de confiance de niveau** α un intervalle aléatoire de la forme $[T(\omega)-A(\omega),T(\omega)+A(\omega)]$ dans lequel "$f(\theta)$ se trouve avec une probabilité au moins égale à α", ce qui veut dire mathématiquement que

$$P_\theta(|T-f(\theta)|\le A)\;\ge\;\alpha,\qquad\forall\theta\in\Theta \tag{17}$$

(ci-dessus, A est une variable aléatoire positive, qu'on choisit souvent égale à une constante).

Exemples:

C (suite). On a un n-échantillon de $\mathcal{N}(\theta,\sigma^2)$ avec σ^2 connu. La loi de $\bar{X}-\theta$ sous P_θ suit la loi $\mathcal{N}(0,\sigma^2/n)$, donc $\sqrt{n}(\bar{X}-\theta)/\sigma$ suit la loi normale centrée réduite. Fixons alors un niveau α, par exemple $\alpha=0,95$. On lit sur la table de la loi normale que

$$P_\theta(|\sqrt{n}(\bar{X}-\theta)/\sigma|>1,96)\;=\;0,05$$

et par suite un intervalle de confiance de niveau $0,95$ pour la moyenne est

$$[\bar{X}-1,96\frac{\sigma}{\sqrt{n}},\bar{X}+1,96\frac{\sigma}{\sqrt{n}}]. \tag{18}$$

A (suite): la loi de $\omega=nT(\omega)$ est $B(\theta,n)$ sous P_θ, donc la variance de T est $\theta(1-\theta)/n$. L'inégalité de Bienaymé-Tchebicheff entraine

$$P_\theta(|T-\theta|>a)\;\le\;\frac{\theta(1-\theta)}{na^2}\;\le\;\frac{1}{4na^2},\qquad\forall\theta\in\Theta.$$

Donc un intervalle de confiance de niveau $0,95$ est donné par

$$[T-\frac{1}{\sqrt{0,2\,n}},T+\frac{1}{\sqrt{0,2\,n}}]. \tag{19}$$

L'inégalité de Bienaymé-Tchebicheff est une approximation assez grossière. On a vu en (4-34) que si n est grand, $(T-\theta)\sqrt{n/\theta(1-\theta)}$ suit approximativement la loi $\mathcal{N}(0,1)$ sous P_θ. On a donc

$$P_\theta(|2\sqrt{n}(T-\theta)| > 1,96) \ \leq \ P_\theta\left(\frac{\sqrt{n}|T-\theta|}{\sqrt{\theta(1-\theta)}} \geq 1,96\right) \ \sim \ 0,05$$

et un intervalle de confiance de niveau $0,95$ est donné par

$$[T - \frac{0,98}{\sqrt{n}}, T + \frac{0,98}{\sqrt{n}}]$$

(qui est plus petit, donc "meilleur", que celui donné par (19), du moins lorsque n est "assez grand" pour que l'approximation normale soit à peu près correcte: dans la pratique, n "assez grand" signifie supérieur à 30).

Considérons maintenant l'exemple G d'un n-échantillon de la loi $\mathcal{N}(m,\sigma^2)$, avec $\theta = (m,\sigma^2)$ inconnu. On veut obtenir un intervalle de confiance pour $f(\theta) = m$. Comme σ^2 est inconnu, on ne peut pas utiliser (18). On peut penser remplacer σ^2 par son (meilleur) estimateur sans biais S^2, donné par (16), mais la variable aléatoire $\sqrt{n}(\bar{X}-m)/S$ (où S est la racine carrée positive de S^2) ne suit pas une loi normale, et nous allons étudier sa loi.

Définition 13. Si $X_1,\ldots,X_n$ est un n-échantillon de $\mathcal{N}(0,1)$, la loi de $U = X_1^2 + \ldots + X_n^2$ s'appelle la **loi du chi-carré à n degrés de liberté**. On la note χ_n^2.

D'après la proposition 3-26 et l'exemple 2 du paragraphe 3-8, on a $\chi_n^2 = \Gamma(n/2,1/2)$. Vu son importance, cette loi est tabulée (bien que pour tout n pair il existe aussi une expression analytique de sa fonction de répartition).

Proposition 14: *Soit $X_1,\ldots,X_n$ un n-échantillon de $\mathcal{N}(m,\sigma^2)$. Alors $\bar{X}$ et S^2 sont indépendantes, $\bar{X}$ suit la loi $\mathcal{N}(m,\sigma^2/n)$, et $(n-1)S^2/\sigma^2$ suit la loi χ_n^2.*

Preuve. L'assertion concernant la loi de $\bar{X}$ a déjà été vue. En faisant le changement de variable $Z_i = (X_i - m)/\sigma$, on remplace $\bar{X}$ par $(\bar{X}-m)/\sigma$ et S^2 par S^2/σ^2: il suffit donc de montrer les autres résultats lorsque $m = 0$ et $\sigma^2 = 1$, ce que nous supposerons par la suite.

Remarquons d'abord que si A est une matrice $n\times n$ orthogonale, les composantes $(Y_1,\ldots,Y_n)$ du vecteur $Y = AX$, où $X = (X_1,\ldots,X_n)$, forment encore un n-échantillon de $\mathcal{N}(0,1)$: cf. la proposition 3-19 et le corollaire 4-12, car AA^{t} est la matrice identité.

Choisissons alors une matrice orthogonale A dont les éléments de la dernière ligne sont tous égaux à $1/\sqrt{n}$. On a $Y_n = (X_1 + \ldots + X_n)/\sqrt{n} = \bar{X}/\sqrt{n}$, et

$$(n-1)S^2 \ = \ \sum_{i=1}^{n} X_i^2 - n\bar{X}^2 \ = \ \sum_{i=1}^{n} Y_i^2 - Y_n^2 \ = \ \sum_{i=1}^{n-1} Y_i^2.$$

Comme les Y_i sont indépendantes de loi $\mathcal{N}(0,1)$, on en déduit que S^2 et Y_n (donc $\bar{X}$) sont indépendantes, et que $(n-1)S^2$ suit la loi χ_{n-1}^2. $\square$

Définition 15. Soit X et Y deux variables aléatoires indépendantes, avec X de loi $\mathcal{N}(0,1)$ et Y de loi χ^2_n. Alors, la loi de $T = X\sqrt{n}/\sqrt{Y}$ s'appelle la **loi de Student à n degrés de liberté**, et on la note t_n.

On peut aisément calculer la densité de t_n. Cette loi est symétrique par rapport à l'origine, et sa fonction de répartition est tabulée pour $n \leq 50$. Pour $n > 50$ cette loi est très voisine de $\mathcal{N}(0,1)$ (Exercice: pourquoi ?).

Proposition 16: *Soit $X_1, \ldots, X_n$ un n-échantillon de $\mathcal{N}(m, \sigma^2)$. Alors $\sqrt{n}(\bar{X} - m)/S$ suit la loi de Student t_{n-1} à $n-1$ degrés de liberté.*

Preuve. Il suffit d'appliquer la définition 15, après avoir remarqué que d'après la proposition 14 les variables aléatoires $\sqrt{n}(\bar{X} - m)/\sigma$ et $(n-1)S^2/\sigma^2$ sont indépendantes et suivent respectivement les lois $\mathcal{N}(0,1)$ et χ^2_{n-1}. $\square$

Exemple G (suite): Cherchons un intervalle de confiance de niveau $\alpha = 0,95$ pour m, lorsque $n = 25$. Si T est une variable de Student t_{24}, on lit dans la table que $P(|T| > 2,06) \sim 0,05$. D'après la proposition 16, l'intervalle cherché est donc

$$[\bar{X} - 2,06\frac{S}{\sqrt{n}}, \bar{X} + 2,06\frac{S}{\sqrt{n}}] \;=\; [\bar{X} - 0,41\ S, \bar{X} + 0,41\ S]. \tag{20}$$

Remarquer que si on avait pris (18) en remplaçant σ^2 par S^2, on aurait obtenu un intervalle trop "optimiste" (i.e., trop petit), ce qui est le genre d'erreur à ne pas faire en statistiques (la différence entre $1,96$ et $2,06$ n'est bien-sûr pas très grande: plus n est grand, et plus la différence entre (18) et (20) est faible).

6 Les modèles linéaires (régression)

Le "modèle linéaire" est un modèle statistique partiellement incomplet, dans lequel on ne spécifie pas entièrement les probabilités P_θ. On peut aussi le considérer comme une modèle non-paramétrique, au sens de l'exemple B.

Définition 17. Un modèle linéaire est la donnée de:

• d paramètres réels **inconnus** $\theta_1, \ldots, \theta_d$. On écrit $\theta = (\theta_1, \ldots, \theta_d)$ le vecteur correspondant de $\mathbb{R}^d$.

• Une matrice $k \times d$ **connue**, soit $A = (a_{ij} : 1 \leq i \leq k, 1 \leq j \leq d)$.

• k variables aléatoires réelles $Y_1, \ldots, Y_k$ centrées (i.e. de moyenne nulle), **non-corrélées** (i.e. elles sont dans $\mathcal{L}^2$ et $\mathrm{cov}(Y_i, Y_j) = 0$ si $i \neq j$), de même variance σ^2 **inconnue** (en général; parfois il se peut qu'on connaisse σ^2). On note $Y = (Y_1, \ldots, Y_k)$ le vecteur aléatoire de composantes Y_i.

• k variables aléatoires réelles $X_1, \ldots, X_k$ qui constituent **l'observation**, et qui sont données par

$$X_j \;=\; \sum_{l=1}^{d} a_{jl}\theta_l + Y_j, \tag{21}$$

ou, en notation matricielle, $X = A\theta + Y$.

Enfin, le problème consiste à **estimer les** θ_l.

On peut se figurer les Y_j comme un "bruit", ou une "erreur d'observation" ou de mesure. Le problème posé n'a réellement de sens que si, dans l'hypothèse où il n'y aurait pas de bruit (i.e. $\sigma^2 = 0$, ou de manière équivalente si toutes les Y_j sont nnulles), on peut retrouver les θ_l à partir des observations: il faut donc que la matrice A **soit de rang** d, et donc en particulier que $k \geq d$.

Exemples:

1) **Régression linéaire:** On a $d = 1$ et $X_j = x_j\theta + Y_j$, où les x_j sont connus et pas tous nuls (cela s'interprète ainsi: on a une fonction linéaire $f(t) = \theta t$ à déterminer; on fait k mesures aux points x_j, et chaque mesure est entachée d'une erreur Y_j).

2) **Régression polynomiale:** Cela correspond au cas particulier suivant de (21):

$$X_j \;=\; \sum_{l=1}^{d} (x_j)^l \theta_l + Y_j. \tag{22}$$

Théorème 18 (de Gauss-Markov): *Supposons la matrice A de rang d, et posons $U = (A^{\mathrm{t}}A)^{-1}A^{\mathrm{t}}$.*

a) La statistique UX est le meilleur estimateur **linéaire sans biais** *du vecteur θ, au sens où pour* **tout** *vecteur $a \in \mathbb{R}^d$, la statistique réelle $a^{\mathrm{t}}UX$ est le meilleur estimateur de $< a, \theta >$ parmi tous les estimateurs qui sont sans biais et qui sont des fonctions affines des observations $X_1, \ldots, X_k$.*

b) La statistique $V = |(AU - I)X|^2$ est un estimateur sans biais de $(k - d)\sigma^2$ (ici, $|.|$ représente la norme euclidienne d'un vecteur de $\mathbb{R}^d$).

c) Si le vecteur aléatoire Y est gaussien (donc les Y_j sont indépendantes: cf. le paragraphe 4-3), alors UX est le meilleur estimateur sans biais de θ (parmi les estimateurs sans biais, linéaires ou non), et V est le meilleur estimateur sans biais de $(k - d)\sigma^2$.

Noter que le modèle statistique ici n'est pas totalement spécifié, puisqu'on n'a pas fixé la loi du vecteur Y (sauf dans (c) ci-dessus). On notera P la probabilité sous-jacente, qui fixe en fait la loi du vecteur Y. Dire que UX est sans biais veut alors dire que $E(UX) = \theta$ lorsque X et Y sont reliés par (21), et ceci pour tout θ **et** pour tout choix de la loi de Y (choix soumis aux conditions de la définition 17).

Preuve. a) Un estimateur linéaire (on devrait d'ailleurs dire "affine") s'écrit $T = BX + Q$, où B est une matrice $d \times k$ et Q est un vecteur de $\mathbb{R}^d$. Comme les Y_j sont centrées et comme $T = BA\theta + BY + Q$, il vient $E(T) = BA\theta + Q$. Donc T est sans biais si et seulement si on $BA\theta + Q = \theta$ pour tout $\theta \in \mathbb{R}^d$, ce qui revient à dire que (avec I_d la matrice identité $d \times d$):

$$BA \;=\; I_d, \qquad Q \;=\; 0. \tag{23}$$

Soit alors $a \in \mathbb{R}^d$. Sous (23), le risque quadratique de la statistique $< a, T >$ pour estimer $< a, \theta >$ est l'espérance de $< a, BX - \theta >^2$. Comme $BX - \theta = BY$ (d'après (23) encore) on a $< a, BX - \theta >^2 = a^t BYY^t B^t a$, et donc le risque quadratique s'écrit

$$R_{<a,T>} \; = \; \sigma^2 a^t BB^t a, \tag{24}$$

puisque la matrice des covariances de Y est $\sigma^2 I_d$.

Soit alors F l'image de $\mathbb{R}^d$ dans $\mathbb{R}^k$ par l'application linéaire associée à la matrice A; soit Π et Π' les matrices $k \times k$ associées aux projections orthogonales dans $\mathbb{R}^k$, respectivement sur F et sur son orthogonal. On a $\Pi + \Pi' = I_k$ et, comme le rang de A égale d, le rang de Π est aussi d, ainsi que la dimension de l'espace F. Il existe donc une matrice U et une seule, qui est $d \times k$ et telle que $U\Pi = U$ et $UA = I_d$. Comme $\Pi = \Pi^t$, une vérification immédiate montre que $U = (A^t A)^{-1} A^t$ (comme A est de rang d, $A^t A$ est bien inversible).

Dire que $BA = I_d$ équivaut à dire que $B(\Pi + \Pi')A = I_d$, donc que $B\Pi A = I_d$ (car $\Pi' A = 0$), donc que $B\Pi = U$. On a alors

$$a^t BB^t a \; = \; a^t (B\Pi + B\Pi')(B\Pi + B\Pi')^t a \; = \; a^t UU^t a + a^t B\Pi' B^t a,$$

et $a^t B\Pi' B^t a \geq 0$ pour tout $a \in \mathbb{R}^d$. Par suite lorsque B varie parmi toutes les matrices vérifiant $B\Pi = U$, le nombre $a^t B\Pi' B^t a$ est minimal, simultanément pour tous les vecteurs a, lorsque $B\Pi' = 0$. Comme $B = B\Pi + B\Pi'$, l'unique matrice vérifiant $B\Pi = U$ et $B\Pi' = 0$ est $B = U$. Par suite $T = UX$ minimise les risques (24), simultanément pour tous les a, et indépendamment de θ, de σ^2, et de la loi (non spécifiée) du vecteur Y, et c'est le seul estimateur affine sans biais jouissant de ces propriétés: on a donc montré la partie (a).

b) On a $AU\Pi A = AUA = A = \Pi A$ (vérification immédiate). Comme l'application linéaire associée à A est bijective de $\mathbb{R}^d$ dans F et que Π est la projection sur F, on déduit de $AU\Pi A = \Pi A$ que $AU\Pi = \Pi$; comme $AU\Pi' = 0$, on a donc $AU = \Pi$. Par suite, comme $\Pi A = A$, la statistique V vaut

$$\begin{aligned} V \; = \; |(\Pi - I_k)X|^2 \; &= \; |(\Pi - I_k)A\theta + (\Pi - I_k)Y|^2 \; = \; |(\Pi - I_k)Y|^2 \\ &= \; |\Pi' Y|^2 \; = \; Y^t \Pi' \Pi'^t Y, \end{aligned}$$

et il vient

$$E(V) \; = \; \sigma^2 \sum_{1 \leq i,j \leq k} \gamma_{ij}^2 \; = \; \sigma^2 (k - d),$$

où γ_{ij} est l'élément (i, j) de la matrice Π'.

c) Lorsque le vecteur Y est gaussien, il en est de même de X et une généralisation immédiate de l'exemple G du paragraphe 5 montre que le modèle (qui est ici complètement spécifié) est exponentiel. Un calcul fastidieux, mais facile, montre que UX est une statistique suffisante, et comme c'est un estimateur sans biais de θ c'est aussi le meilleur estimateur sans biais. De même, V est une fonction de UX, donc c'est le meilleur estimateur sans biais de $(k - d)\sigma^2$. □

Exemple: Considérons le modèle de régression linéaire $X_j = x_j \theta + Y_j$. En calculant U (ce qui est facile ici, car $d = 1$) on voit que le meilleur estimateur linaire sans biais de θ est

$$UX \; = \; \frac{\sum_{j=1}^k x_j X_j}{\sum_{j=1}^k (x_j)^2}.$$

On remarquera que UX est aussi la pente a de la droite $y = ax$ qui "passe le mieux au milieu des observations" dans le sens où a minimise $\sum_j (X_j - ax_j)^2$: cette droite est appelée **droite des moindres carrés**.

7 Autres méthodes d'estimation

A titre de compléments, nous indiquons rapidement ci-dessous deux autres méthodes d'estimation qui sont couramment utilisées dans la pratique.

Pour la première, nous supposons que Θ est un ouvert de $I\!R^p$ pour un certain entier p. La difficulté essentielle dans le paragraphe 2 est que si S et T sont deux estimateurs de $f(\theta)$, leurs risques quadratiques R_S et R_T ne sont en général pas comparables. D'où l'idée de se donner une probabilité μ sur Θ, dite **probabilité a priori**, et de définir le **risque quadratique bayésien**

$$\rho_T \;=\; \int R_T(\theta)\mu(d\theta) \;=\; \int E_\theta[(T-f(\theta))^2]\mu(d\theta) \tag{25}$$

relatif à la probabilité a priori μ.

Définition 19. On appelle **estimateur bayésien** de $f(\theta)$ (pour la probabilité a priori μ) un estimateur T tel que $\rho_T \leq \rho_S$ pour toute autre statistique S.

L'intérêt de cette notion est que, avec μ fixée, les nombres ρ_T et ρ_S sont forcément comparables (alors que les fonctions R_T et R_S ne le sont en général pas). En outre, comme nous allons le voir, on peut calculer les estimateurs bayésiens dans de nombreux cas. L'inconvénient est bien-sûr le caractère arbitraire de la probabilité a priori μ (si on change de probabilité μ, on change en général d'estimateur bayésien).

Nous allons supposer que

$$\Omega = I\!R^d, \qquad P_\theta \text{ admet une densité } h_\theta(x) = h(\theta,x). \tag{26}$$

Soit T un estimateur de $f(\theta)$. On a

$$\rho_T \;=\; \int \mu(d\theta)\int [T(x)-f(\theta)]^2 h(\theta,x)dx \;=\; \int dx \int \mu(d\theta)h(\theta,x)[T(x)-f(\theta)]^2$$

(on supposera l'interversion des deux intégrales permise). Nous devons choisir T de sorte que cette expression soit minimale, et il n'est pas difficile de vérifier que c'est le cas si

$$T(x) \;=\; \frac{\int f(\theta)h(\theta,x)\mu(d\theta)}{\int h(\theta,x)\mu(d\theta)} \tag{27}$$

($=0$ si le dénominateur est nul), pourvu bien-sûr que les intégrales ci-dessus aient un sens. Autrement dit, (27) définit un **estimateur bayésien** de $f(\theta)$.

Pour décrire la sconde méthode, qui est de nature entièrement différente, nous supposons encore (26). Cette méthode permet d'estimer θ directement, quelle que soit la nature de l'ensemble Θ (contrairement aux méthodes précédentes, qui ne permettent que d'estimer une fonction réelle, ou au plus vectorielle, de θ).

Définition 20. Supposons (26). Un estimateur T à valeurs dans Θ est appelé **estimateur du maximum de vraisemblance** si

$$h(T(x),x) \;=\; \sup_{\theta\in\Theta} h(\theta,x) \qquad \forall x\in\Omega = I\!R^d \tag{28}$$

(la fonction h est quelquefois appelée "fonction de vraisemblance").

Bien entendu, un tel estimateur n'existe pas toujours (la fonction $\theta \mapsto h(\theta, x)$ pouvant ne pas atteindre son supremum), et il peut aussi être multiple (pire: il peut ne pas exister pour certaines valeurs de x, et être multiple pour d'autres!). Cependant dans la plupart des cas concrets cet estimateur existe, est unique, et est en plus raisonnablement facile à calculer.

Il n'y a pas vraiment de justification théorique à cette méthode, si ce n'est asymptotiquement: sous des hypothèses très générales, l'estimateur du maximum de vraisemblance d'un n-échantillon converge vers la vraie valeur du paramètre lorsque $n \to \infty$.

Signalons aussi une qualité importante de cet estimateur: il est **invariant par changement de paramètre**. Si par exemple $\Theta = I\!R_+$ on peut, de manière équivalente, paramètrer le modèle par θ^2 ou $\sqrt{\theta}$ au lieu de θ. L'estimateur du maximum de vraisemblance pour θ^2 est évidemment le carré de celui pour θ; par contre si T est le meilleur estimateur sans biais de θ (à supposer qu'il existe), son carré T^2 n'est pas le meilleur estimateur sans biais de θ^2 (en particulier, il est biaisé).

Exemples:

C (suite): On observe un n-échantillon de $\mathcal{N}(\theta, \sigma^2)$ avec σ^2 connu, et $\Theta = I\!R$. On a (26) avec h donné par (8). Il est immédiat de vérifier que dans ce cas la moyenne empirique est l'estimateur du maximum de vraisemblance.

G (suite): On observe un n-échantillon de $\mathcal{N}(m, \sigma^2)$ avec $\theta = (m, \sigma^2)$ inconnu. On a encore (26) avec h donné par (8). Il est plus aisé de prendre pour paramètres le couple (m, ν), avec $\nu = 1/\sigma^2$, de sorte qu'on a à maximiser la fonction suivante de (m, ν) pour chaque $x = (x_1, \dots, x_n)$ fixé:

$$k(m, \nu; x) = \left(\frac{\nu}{2\pi}\right)^{n/2} \exp - \sum_{i=1}^{n} \frac{\nu}{2}(x_i - m)^2.$$

En dérivant, il est facile de voir que le maximum est atteint pour

$$m = \frac{1}{n}(x_1 + \ldots + x_n), \qquad \nu = \frac{n}{(x_1 - m)^2 + \ldots + (x_n - m)^2}.$$

En d'autres termes, l'estimateur du maximum de vraisemblance pour le couple (m, σ^2) est le couple

$$\left(\bar{X}, \frac{n-1}{n} S^2\right). \tag{29}$$

(cf. (16)). Pour n grand, il est donc peu différent du meilleur estimateur sans biais.

CHAPITRE 6

Statistiques: les tests

1 Introduction

Soit $(\Omega, \mathcal{A}, (P_\theta)_{\theta\in\Theta})$ un modèle statistique. l'ensemble Θ est ici divisé en une partie Θ_0 et son complémentaire Θ_1. L'objectif d'un test est, au vu de l'observation ω, de "décider" si la vraie valeur de θ se trouve dans Θ_0 ou dans Θ_1.

Exemple: Reprenons le premier exemple du chapitre précédent. Soit θ_0 la valeur limite de la proportion de pièces défectueuses qui est acceptable. On veut décider, au vu du nombre X de pièces défectueuses observées dans un échantillon de n pièces, si $\theta > \theta_0$ ou non.

Le problème ci-dessus, comme la plupart des problèmes pratiques de ce type, est essentiellement dissymétrique: on veut être "raisonnablement sûr" que $\theta \leq \theta_0$; autrement dit, on veut rejeter l'hypothèse "$\theta > \theta_0$" avec une "erreur" faible si elle est vraie (parce que si la machine est mal réglée, les clients vont refuser les lots de pièces qui contiendront "en moyenne" trop de pièces défectueuses); par contre si la vraie valeur est $\theta \leq \theta_0$ et si on décide à tort qu'elle est plus grande, ce n'est pas très grave: on en est quitte pour un réglage supplémentaire - inutile - de la machine.

Définition 1. a) Dans un problème de test, on veut **tester l'hypothèse** H_0 selon laquelle $\theta \in \Theta_0$, **contre l'alternative** H_1 selon laquelle $\theta \in \Theta_1$ (la terminologie exprime bien le caractère dissymétrique).

b) La **région critique** est la partie (ou, l'événement) D de Ω sur lequel on rejette l'hypothèse H_0: si $\omega \in D$, on dit que H_1 est satisfaite, et H_0 dans le cas contraire. On dit aussi que D est un "test".

c) Si $\theta \in \Theta_0$, le nombre $P_\theta(D)$ (probabilité de rejeter l'hypothèse H_0 alors qu'elle est vraie) s'appelle **l'erreur de première espèce**.

d) Si $\theta \in \Theta_1$, le nombre $1 - P_\theta(D) = P_\theta(D^c)$ (probabilité d'accepter l'hypothèse H_0 alors qu'elle est fausse) s'appelle **l'erreur de seconde espèce**.

e) Le nombre $\alpha = \sup_{\theta_0\in\Theta_0} P_\theta(D)$ est le **niveau** du test, ou de la région critique.

f) La fonction $\theta \mapsto P_\theta(D)$ s'appelle la **fonction puissance** du test.

Il s'agit donc de "construire un test", ce qui veut dire trouver une région critique, qui minimise autant que possible les erreurs de première et de seconde espèce. Il s'agit d'un problème mathématique encore plus difficile que trouver des estimateurs optimaux: en effet, très souvent Θ est une partie de $I\!R^d$, et pour n'importe quelle région critique D la fonction puissance est continue; or on cherche à la rendre aussi proche que possible de 1 sur Θ_1 et de 0 sur Θ_0,: ceci est manisfestement contradictoire, en général, au voisinage de la frontière entre Θ_0 et Θ_1.

En fait, tirant parti du caractère dissymétrique, les tests sont construits de la manière suivante:

1) On fixe une borne supérieure au niveau α, en général $0,1$ ou $0,05$ ou $0,01$.

2) Parmi toutes les régions critiques de niveau α (ou $\leq \alpha$), on cherche à maximiser la fonction puissance sur Θ_1, c'est-à-dire à minimiser les erreurs de seconde espèce (avec les mêmes problèmes que pour minimiser les risques des estimateurs: les fonctions puissance de deux régions critiques ne sont en général pas comparables).

Implicitement, cela veut dire qu'on considère comme plus grave une erreur de première espèce qu'une erreur de seconde espèce: les premières sont majorées uniformément par le "petit" nombre α, tandis que les secondes sont souvent proches de $1-\alpha$ aux points de Θ_1 qui sont "proches" de la frontière avec Θ_0.

Cette méthode conduit au critère de qualité suivant:

Définition 2. Soit D un test de niveau α.

a) Un autre test D' est dit **plus puissant** que D s'il est de niveau $\alpha' \leq \alpha$ et si $P_\theta(D') \geq P_\theta(D)$ pour tout $\theta \in \Theta_1$.

b) Le test D est **uniformément plus puissant** (en abrégé: UPP) s'il n'existe pas de test D' plus puissant que D et vérifiant en outre $P_\theta(D') > P_\theta(D)$ pour une valeur θ au moins dans Θ_1.

2 Tests sur la loi normale

Nous n'allons pas exposer la théorie générale des tests, ce qui dépasserait de beaucoup les objectifs de ce cours, mais nous contenter de donner des exemples, les plus significatifs, et la plupart du temps sans démonstration. De toutes façons, il existe peu de cas où l'on connaisse des tests UPP, et la plupart du temps on se contente de tests "plausibles", qu'on valide empiriquement en faisant des simulations. Cependant, **le** point essentiel est que le niveau d'un test soit calculable; pour le reste, on espère que la fonction puissance est "suffisamment" bonne.

Signalons toutefois que si T est une **statistique suffisante**, pour tout test D il existe un test D' qui est meilleur que D (rappelons que cela signifie en fait "au moins aussi bon"), et qui est de la forme $D' = \{T \in A\}$ pour un ensemble A convenable: c'est l'analogue de la proposition 5-8.

Dans ce paragraphe, on donne une liste de tests sur la loi normale. On part d'un n-échantillon $X_1, \ldots, X_n$ de $\mathcal{N}(m, \sigma^2)$. On note $P_{m,\sigma}$ la probabilité correspondante (que m

et/ou σ^2 soient connus ou inconnus). Rappelons qu'on pose

$$\bar{X} \;=\; \frac{1}{n}(X_1 + \ldots + X_n), \qquad S^2 \;=\; \frac{1}{n-1}\Big((X_1 - \bar{X})^2 + \ldots + (X_n - \bar{X})^2\Big), \tag{1}$$

$$\Sigma^2 \;=\; \frac{1}{n}\Big((X_1 - m)^2 + \ldots + (X_n - m)^2\Big) \qquad \text{si } m \text{ est connu.} \tag{2}$$

Rappelons aussi que:

- Si m est inconnu et σ^2 connu, $\bar{X}$ est suffisante.
- Si m est connu et σ^2 inconnu, Σ^2 est suffisante.
- Si m et σ^2 sont inconnus, $(\bar{X}, S^2)$ est suffisante.

1) σ^2 connu, test de $m \leq m_0$ contre $m > m_0$: On montre que les tests de région critique

$$D \;=\; \{\bar{X} \geq a\} \tag{3}$$

sont UPP. Le niveau du test (3) est $\sup_{m \leq m_0} P_{m,\sigma}(\bar{X} > a)$, qui vaut clairement $P_{m_0,\sigma}(\bar{X} > a)$. Comme sous $P_{m_0,\sigma}$ la variable aléatoire $\sqrt{\bar{X} - m_0)}/\sigma$ suit la loi $\mathcal{N}(0,1)$, ce niveau se calcule facilement à partir de la table de la fonction de répartition de cette loi $\mathcal{N}(0,1)$.

Pour des raisons évidentes de symétrie, les tests $\{\bar{X} \leq a\}$ sont UPP pour tester $m \geq m_0$ contre $m < m_0$.

2) σ^2 connu, test de $m = m_0$ contre $m \neq m_0$: On montre que le test de région critique

$$D \;=\; \{|\bar{X} - m_0| \geq a\} \tag{4}$$

est UPP, de niveau $P_{m_0,\sigma}(|\bar{X} - m_0| > a)$, qui se calcule facilement à partir de la table de la fonction de répartition de la loi $\mathcal{N}(0,1)$.

3) m et σ^2 inconnus, test de $m \leq m_0$ contre $m > m_0$: On utilise la statistique $T_{m_0} = \sqrt{n}(\bar{X} - m_0)/S$, dont on sait (proposition 5-16) qu'elle suit une loi de Student t_{n-1} sous $P_{m_0,\sigma}$. On montre alors que les tests de région critique

$$D \;=\; \{T_{m_0} \geq a\} \tag{5}$$

sont UPP. Le niveau du test (5) est $P_{m_0,\sigma}(T_{m_0} > a)$ (rappelons que cette quantité ne dépend pas de σ), et il se calcule à partir de la table de la fonction de répartition de la loi t_{n-1}.

Pour des raisons évidentes de symétrie, les tests $\{T_{m_0} \leq a\}$ sont UPP pour tester $m \geq m_0$ contre $m < m_0$.

Exemple: Supposons $n = 20$. On veut tester $m \leq 0$ contre $m > 0$, avec un niveau $\alpha = 0,05$. En utilisant la table de la loi de Student t_{19}, on voit que $a = 1,64$ dans (5). On rejette donc l'hypothèse si la quantité (calculable à partir des observations) $T_0 = \sqrt{20}\bar{X}/S$ vérifie $T_0 \geq 1,64$.

4) m **et** σ^2 **inconnus, test de** $m = m_0$ **contre** $m \neq m_0$**:** On utilise la même statistique T_{m_0}, et les tests de région critique

$$D \;=\; \{|T_{m_0}| \geq a\}, \tag{6}$$

bien que ces tests ne soient pas UPP. Le niveau du test (6) est $P_{m_0,\sigma}(|T_{m_0}| > a)$ (qui, encore une fois, ne dépend pas de σ et se calcule à partir de la table de la loi t_{n-1}.

5) m **connu, test de** $\sigma \geq \sigma_0$ **contre** $\sigma < \sigma_0$**:** On peut maintenant utiliser la statistique Σ^2, et les tests de région critique

$$D \;=\; \{\Sigma^2 \leq a\} \tag{7}$$

sont UPP. De plus sous P_{m,σ_0} la variable aléatoire Σ^2/σ_0^2 suit d'après la proposition 5-14 la loi du chi-carré χ^2_n, indépendamment de la valeur de m. Donc le niveau du test (7) est $P_{m,\sigma_0}(\Sigma^2 > a)$ (quantité indépendante de m) et est donné par les tables de la loi χ^2_n.

6) m **inconnu, test de** $\sigma \geq \sigma_0$ **contre** $\sigma < \sigma_0$**:** On peut montrer que les tests de région critique

$$D \;=\; \{S^2 \leq a\} \tag{8}$$

sont UPP. De plus sous P_{m,σ_0} la variable aléatoire $(n-1)S^2/\sigma_0^2$ suit la loi χ^2_{n-1}, indépendamment de la valeur de m. Donc le niveau du test (7) est $P_{m,\sigma_0}(S^2 > a)$ (quantité indépendante de m) et est donné par les tables de loi χ^2_{n-1}.

Remarque importante: Dans les tests ci-dessus, il faut faire très attention aux paramètres inconnus. Par exemple dans le cas (3) on pourrait être tenté d'estimer plus ou moins grossièrement σ^2, puis d'appliquer les tests décrits en (1): cela conduit à des erreurs **graves**, car le niveau est alors sous-évalué (ou, ce qui revient au même, à niveau donné on donne une région critique qui est trop grande, donc l'erreur de première espèce commise effectivement est plus grande que celle qu'on calcule; or, c'est l'erreur de première espèce qu'on vise avant tout à contrôler précisément).

3 Tests d'adéquation

Dans ce paragraphe, on est essentiellement dans un cadre "non-paramétrique": l'espace Θ est très grand, c'est en fait l'ensemble de toutes les probabilités sur un espace donné E.

Plus précisément, on considère un n-échantillon $X_1, \ldots, X_n$ de loi μ sur une ensemble E (en général, un ensemble fini ou dénombrable, ou $E = \mathbb{R}^d$). On veut tester

H_0: On a $\mu = \mu_0$, où μ_0 est une loi donnée,

contre

H_1: On a $\mu \neq \mu_0$.

Nous allons décrire le **test du** χ^2. En premier lieu on choisit une partition finie $E_1, \ldots, E_q$ de l'espace E, et on pose

$$p_j \;=\; \mu_0(E_j). \tag{9}$$

On suppose que $p_j > 0$ pour tout j (dans la pratique, on choisit la partition de sorte que les p_j soient le plus proche possible de $1/q$). Ensuite, on note N_j^n le nombre de variables aléatoires X_i qui tombent dans E_j (on a donc $N_1^n + \ldots + N_q^n = n$), et on pose

$$T_n \;=\; \sum_{j=1}^{q} \frac{1}{np_j}(N_j^n - np_j)^2. \tag{10}$$

L'idée du test du χ^2 est la suivante: si $\mu = \mu_0$, on a convergence presque sûre de N_j^n/n vers p_j lorque $n \to \infty$ d'après la loi ds grands nombres; on peut donc espérer que T_n ne deviendra pas trop grand pour $n \to \infty$ (on verra qu'en fait T_n converge en loi vers une variable aléatoire finie). En revanche si $\mu \neq \mu_0$ on a $N_j^n/n \to p'_j = \mu(E_j)$ p.s.; par suite si $p'_j \neq p_j$ pour au moins une valeur de j, les variables aléatoires T_n tendent vers l'infini. Cela conduit à choisir une région critique de la forme

$$D \;=\; \{T_n \geq a\}. \tag{11}$$

Ces régions critiques ne sont nullement UPP, et on ne connait d'ailleurs pas de test UPP pour ce problème. Mais elles semblent raisonnables et sont universellement employées.

Il nous reste, ce qui est essentiel, à déterminer le niveau du test (11). Ce niveau est $P_{\mu_0}(D)$, où P_{μ_0} est la probabilité sur notre espace d'états lorsque la loi des X_i est μ_0. Comme μ_0 est connu, on peut en principe calculer ce niveau: en effet, sous P_{μ_0}, la variable aléatoire q-dimensionnelle $S_n = (N_1^n, \ldots, N_q^n)$ suit une **loi multinomiale**, donnée par

$$P_{\mu_0}(N_1^n = i_1, \ldots, N_q^n = i_q) \;=\; n! \prod_{j=1}^{q} \frac{(p_j)^{i_j}}{i_j!} \qquad \text{si } \; i_1 + \ldots + i_q = n. \tag{12}$$

On peut alors en déduire la loi de T_n et calculer $P_{\mu_0}(T_n \geq a)$.

Mais les calculs effectifs sont compliqués, et numériquement impossibles dès que n est grand. On préfère utiliser ce qui suit:

Proposition 3: *Si $\mu = \mu_0$, les T_n convergent en loi (quand $n \to \infty$) vers une variable aléatoire admettant la loi χ^2_{q-1} à q degrés de liberté.*

Preuve. Soit S_n comme ci-dessus. Soit Y_j la variable q-dimensionnelle dont la $i^{\text{ème}}$ composante vaut 1 si $X_j \in E_i$ et 0 sinon. Les variables Y_j sont indépendantes, de même loi, et $S_n = Y_1 + \ldots + Y_n$.

Soit m le vecteur espérance de Y_1 et $C = (c_{ij})_{1 \leq i,j \leq q}$ sa matrice de covariance. Comme la probabilité pour que $Y_1 = (0, \ldots, 0, 1, 0, \ldots, 0)$, avec un “1” à la $i^{\text{ème}}$ place et des “0” ailleurs, vaut p_i, il est facile de voir que

$$m_i = p_i, \qquad c_{ij} = \begin{cases} p_i(1-p_i) & \text{si } \; i = j \\ -p_i p_j & \text{sinon.} \end{cases}$$

De plus, d'après la version q-dimensionnelle du théorème central limite 4-27, les variables aléatoires $V_n = (S_n - nm)/\sqrt{n}$ convergent en loi vers un vecteur gaussien centré de covariance C. Si $V_{n,i}$ désigne la $i^{\text{ème}}$ composante de V_n, un calcul immédiat fournit

$$T_n \;=\; \sum_{i=1}^{q} \frac{(V_{n,j})^2}{p_j}.$$

Par suite si $U = (U_1, \ldots, U_q)$ est un vecteur gaussien centré de covariance C, les variables aléatoires T_n convergent en loi vers

$$T \;=\; \sum_{i=1}^{q} \frac{U_j^2}{p_j}.$$

Soit alors A une transformation orthogonale sur $\mathbb{R}^q$ telle que le vecteur unitaire de coordonnées $(\sqrt{p_1}, \ldots, \sqrt{p_q})$ soit transformé en $(0, \ldots, 0, 1)$. Soit $W = AU'$, où U' est le vecteur aléatoire de composantes $U_i/\sqrt{p_i}$. La covariance C' de U' a pour éléments $c'_{ii} = 1 - p_i$ et $c'_{ij} = -\sqrt{p_i p_j}$ si $i \neq j$, et la covariance de W (qui est centré) est $K = AC'A^{\mathrm{t}}$. Un calcul simple montre que ses éléments sont $k_{ij} = 0$ si $i \neq j$ ou si $i = j = q$, et $k_{ii} = 1$ si $1 \leq i \leq q-1$. Autrement dit on a $W_q = 0$ p.s., et les $(W_i)_{1 \leq i \leq q-1}$ sont indépendantes de loi $\mathcal{N}(0,1)$. Enfin, comme A est orthogonale on a $|W|^2 = |U'|^2$, de sorte que

$$T \;=\; \sum_{i=1}^{q-1} (W_i)^2 \qquad \text{p.s.}$$

Par suite la loi de T est la loi χ^2_{q-1} par la définition 5-13. $\square$

Pratiquement, on opère ainsi: on choisit les E_j de sorte que pour chaque j le produit np_j soit “assez grand” (disons, ≥ 10), et les p_j aussi voisins que possible les uns des autres. Dans ce cas l'approximation de la loi de T_n par χ^2_{q-1} est bonne. Le niveau étant fixé, on lit alors sur les tables de loi du chi-carré la valeur de a dans (11).

CHAPITRE 7

Le processus de Poisson

1 Introduction

Les probabilités ne servent pas seulement à modéliser des variables aléatoires "individuelles". L'une des applications les plus importantes de la théorie consiste à modéliser les phénomènes aléatoires qui dépendent du temps: on dit qu'on a alors à faire à un **processus stochastique**. On appelle ainsi une famille de variables aléatoires $(X_t)_{t\in T}$ indicée par un ensemble T qui représente le temps: soit le processus est à temps "discret", i.e. $T = \mathbb{N}$ ou $T = \mathbb{Z}$, soit il est à temps "continu" et alors $T = \mathbb{R}_+$ ou $T = \mathbb{R}$. A l'inverse des variables aléatoires qui, quand on en considère une suite finie ou infinie, sont très souvent indépendantes, pour un processus stochastique les variables aléatoires X_t pour différentes valeurs du temps t sont en général fortement dépendantes.

Dans ce chapitre nous allons considérer un processus à temps continu, avec $T = \mathbb{R}_+$. Il s'agit du *processus de Poisson*, qui est l'un des processus les plus simples, et qui modélise une vaste gamme d'applications concrètes de manière extrêmement précise.

Il s'agit de ce qu'on appelle un "processus de comptage", ou une "répartition ponctuelle": un point de vue consiste à imaginer qu'une succession d'événements se produit au cours du temps, et on observe leurs instants d'occurrence, qu'on note $T_1, \ldots, T_n, \ldots$; par hypothèse ces temps sont strictement positifs, et la suite (T_n) est strictement croissante. On "complète" cette suite en posant $T_0 = 0$ (qui n'est pas le temps d'arrivée d'un événement). Un autre point de vue, équivalent sur le plan mathématique, consiste à considérer pour chaque temps $t \geq 0$ le nombre N_t d'événements qui se sont produits sur l'intervalle $[0, t]$: la fonction $t \mapsto N_t$ est donc nulle en 0, croissante, continue à droite, à valeurs dans $\mathbb{N} \cup \{+\infty\}$, et avec des "sauts" d'amplitude 1. On peut facilement passer de la description en termes des (T_n) à celle en termes des (N_t) et vice-versa, via les formules:

$$\left.\begin{aligned} N_t &= n \qquad \text{si } T_n \leq t < T_{n+1} \\ T_n &= \inf(t : N_t = n) \end{aligned}\right\} \tag{1}$$

Une troisième manière, encore équivalente, de décrire le phénomène consiste à introduire les intervalles de temps $S_1, \ldots, S_n, \ldots$ qui séparent deux événements successifs. Ces temps se calculent à partir des T_n, et vice-versa, à partir des formules:

$$\left.\begin{array}{rcl} S_n & = & T_n - T_{n-1} \qquad \text{si } \ n \geq 1 \\ T_0 & = & 0, \qquad T_n \ = \ S_1 + \ldots + S_n \qquad \text{si } \ n \geq 1. \end{array}\right\} \qquad (2)$$

Jusqu'à présent nous n'avons pas fait intervenir de probabilité, et ce qui précède décrit quantitativement toute succession d'événements "discrets" au cours du temps. Dorénavant, nous supposerons que les quantités ci-dessus sont aléatoires, c'est-à-dire que les T_n, les S_n et les N_t sont des variables aléatoires définies sur le même espace d'états Ω, et bien-sûr liées par les formules (1) et (2).

Pour compléter la modélisation, il faut enfin décrire la probabilité ou, ce qui revient au même, la loi "jointe" de la famille de variables aléatoires $(T_n)_{n\geq 1}$, ou de la famille $(S_n)_{n\geq 1}$, ou de la famille $(N_t)_{t\geq 0}$: étant données les formules ci-dessus, ces trois points de vue sont équivalents. C'est la description de cette probabilité qui constitue l'essence même de la modélisation du phénomène.

Par exemple, on peut imaginer que les T_n pour $n \geq 1$ sont les temps de passage des autobus à un arrêt donné, et les S_n sont donc les intervalles de temps entre deux passages successifs. Si les autobus arrivent de manière rigoureusement périodique, on prendra la probabilité pour laquelle les S_n sont toutes égales à la période s. Si au contraire ils arrivent de manière plus ou moins aléatoire, il est naturel de considérer la suite (S_n) comme une suite de variables aléatoires dont la loi est fixée par les conditions de trafic, la disponibilité des chauffeurs, etc...

2 Construction du processus de Poisson

Nous nous plaçons dans le cadre décrit ci-dessus. On dit que le processus (N_t) est un **processus de Poisson**, ou de manière équivalente que la "répartition de points" (T_n) est une **répartition de Poisson**, si les variables aléatoires $(S_n)_{n\geq 1}$ sont indépendantes et de même loi exponentielle. Le paramètre $\lambda > 0$ des ces lois est le **paramètre** du processus de Poisson.

Ce modèle ne décrit pas très bien les arrivées d'autobus à un arrêt; il décrit en revanche de manière extrêmement précise les instants successifs d'appels à un standard téléphonique.

Supposons donc que notre processus (N_t) soit un processus de Poisson de paramètre λ. En vertu de (1) et (2), il doit donc être possible de déterminer la loi des familles (T_n) et (N_t). Pour les T_n, c'est assez facile. D'abord, en vertu de la proposition 3-26, T_n suit la loi gamma $\Gamma(n, \lambda)$. Cela est bien-sûr insuffisant, car les T_n ne sont pas indépendantes (puisque, par exemple, $T_{n+1} > T_n$), et nous devons aussi donner la loi du vecteur aléatoire $(T_1, \ldots, T_n)$. Pour cela, posons

$$\left.\begin{array}{rcl} A_n & = & \{(x_1, \ldots, x_n) \in I\!R^n : 0 < x_1 < \ldots < x_n\}, \\ A_{n,t} & = & \{(x_1, \ldots, x_n) \in I\!R^n : 0 < x_1 < \ldots < x_n \leq t\} \end{array}\right\} \qquad (3)$$

Proposition 1: *Pour chaque entier $n \geq 1$ le vecteur aléatoire $(T_1, \ldots, T_n)$ admet la densité suivante sur $I\!R^n$:*

$$f_n(x_1, \ldots, x_n) \ = \ \lambda^n \ 1_{A_n}(x_1, \ldots, x_n) \ e^{-\lambda x_n}. \qquad (4)$$

Preuve. Comme les S_i sont indépendantes, de densité $\lambda e^{-\lambda y} 1_{\mathbb{R}_+}(y)$, pour toute fonction g sur $\mathbb{R}^n$ il vient

$$\begin{aligned} E[g(T_1,\ldots,T_n)] &= E[g(S_1, S_1+S_2,\ldots,S_1+\ldots+S_n)] \\ &= \lambda^n \int_{\mathbb{R}_+^n} g(y_1,\ldots,y_1+\ldots+y_n) e^{-\lambda(y_1+\ldots+y_n)} dy_1\ldots dy_n, \end{aligned}$$

et le changement de variables $x_1 = y_1$, $x_2 = y_1 + y_2$, …, $x_n = y_1 + \ldots + y_n$, de jacobien 1, montre que l'expression précédente vaut $\int g(x) f_n(x) dx$ si f_n est donnée par (4), d'où le résultat. $\square$

Pour les (N_t), les choses sont plus compliquées et la "loi du processus" (N_t) sera déterminée dans le paragraphe suivant. Il est cependant facile de trouver la loi de chaque variable N_t:

Proposition 2: *Pour chaque $t > 0$ la variable aléatoire N_t suit une loi de Poisson de paramètre λt.*

Preuve. Comme $\{N_t = n\} = \{T_n \leq t < T_{n+1}\}$, il vient d'après (4):

$$\begin{aligned} P(N_t = n) &= \lambda^{n+1} \int_{A_{n+1}} 1_{\{x_n \leq t < x_{n+1}\}} e^{-\lambda x_{n+1}}\, dx_1 \ldots dx_{n+1} \\ &= \lambda^n \int_{A_{n,t}} dx_1 \ldots dx_n \int_t^\infty \lambda\, e^{-\lambda z} dz \;=\; \lambda^n\, e^{-\lambda t}\, \mu(A_{n,t}), \end{aligned}$$

où $\mu(A_{n,t})$ est le volume dans $\mathbb{R}^n$ de l'ensemble $A_{n,t}$. On remarque facilement que $\mu(A_{n,t}) = t^n/n!$, car le cube de $\mathbb{R}^n$ de coté t peut être divisé en $n!$ parties de même volume, à savoir les ensembles où $0 < x_{\sigma_1} < \ldots < x_{\sigma_n} < t$ pour toutes les permutations $\sigma_1, \ldots, \sigma_n$ des entiers $\{1, \ldots, n\}$, et $A_{n,t}$ est l'ensemble ci-dessus correspondant à la permutation $\sigma_i = i$. Par suite $P(N_t = n) = e^{-\lambda t}(\lambda t)^n/n!$, d'où le résultat.

(On aurait aussi pu utiliser la propriété $P(N_t = n) = P(T_n \leq t) - P(T_{n+1} \leq t)$ et la densité (3-37) des lois gamma). $\square$

3 Les accroissements du processus de Poisson

Généralisant la définition 3-20, on dira qu'une famille quelconque de variables aléatoires est **indépendante** si toute sous-famille finie est indépendante.

Proposition 3: *Soit $t \in \mathbb{R}_+$ et $N'_s = N_{t+s} - N_t$. Alors le processus $(N'_s)_{s \in \mathbb{R}_+}$ est un processus de Poisson de paramètre λ, indépendant des variables $(N_r)_{r \in [0,t]}$.*

En particulier, $N_{t+s} - N_t$ est indépendant des $(N_r)_{r\in[0,t]}$, et de même loi que N_s: on dit que (N_t) est un **processus à accroissements indépendants et stationnaires**.

Preuve. Les N'_s sont définis à partir des formules (1) et (2), à l'aide de la suite (S'_n) construite ainsi:

$$S'_1 \;=\; T_{n+1} - t, \qquad S'_i \;=\; S_{n+i} \;\text{ si }\; i \geq 2,$$

sur l'ensemble $\{N_t = n\}$. Il suffit donc de montrer que les (S'_n) sont de loi exponentielle de paramètre λ, indépendantes entre elles, et indépendantes des $(N_r)_{0\leq r\leq t}$. Cela revient à montrer que pour tout entier q, tous réels t_i avec $0 = t_0 < t_1 < \ldots < t_q = t$, tous entiers n et n_i, et enfin pour toute fonction positive bornée f sur $I\!R^q$, on a

$$E[f(S'_1,\ldots,S'_n)1_{\{N_{t_1}=n_1,\ldots,N_{t_q}=n_q\}}] \;=\; E[f(S'_1,\ldots,S'_n)]\; P(N_{t_1}=n_1,\ldots,N_{t_q}=n_q).$$

Notons α le premier membre ci-dessus, et β et γ les deux facteurs du second membre. On a clairement

$$\beta \;=\; \lambda^n \int_{I\!R^n_+} e^{-\lambda(x_1+\ldots+x_n)}\; f(x_1,\ldots,x_n)dx_1\ldots dx_n. \tag{5}$$

Si maintenant

$$B \;=\; \left(\cap_{i=1}^{q-1}\{x_1+\ldots+x_{n_i} \leq t_i < x_1+\ldots+x_{n_i+1}\}\right) \cap \{x_1+\ldots+x_{n_q} \leq t_q\}$$

et si pour simplifier on pose $p = n_q$ et $m = n+p$, on a

$$\{N_{t_1}=n_1,\ldots,N_{t_q}=n_q\} \;=\; \{(S_1,\ldots,S_p)\in B\}\cap\{S_1+\ldots+S_{p+1}>t\},$$

de sorte que

$$\begin{aligned}
\gamma &= \lambda^{p+1}\int_{I\!R^{p+1}_+} e^{-\lambda(x_1+\ldots+x_{p+1})}1_B(x_1,\ldots,x_p)1_{\{x_1+\ldots+x_{p+1}>t\}}dx_1\ldots dx_{p+1}\\
&= \lambda^p\int_B e^{-\lambda(x_1+\ldots+x_p)}dx_1\ldots dx_p\int_{t-x_1-\ldots-x_p}^{\infty}\lambda e^{-\lambda x_{p+1}}dx_{p+1}\\
&= \lambda^p\; e^{-\lambda t}\int_B dx_1\ldots dx_p. \qquad (6)
\end{aligned}$$

Enfin comme $S'_1 = S_1+\ldots+S_{p+1}-t$ sur $\{N_t = p\}$, on a aussi

$$\begin{aligned}
\alpha \;=\; \lambda^m \int_{I\!R^{m+1}_+} & e^{-\lambda(x_1+\ldots+x_m)}1_B(x_1,\ldots,x_p)1_{\{x_1+\ldots+x_{p+1}>t\}}\\
& f(x_1+\ldots+x_{p+1}-t,x_{p+2},\ldots,x_m)dx_1\ldots dx_m.
\end{aligned}$$

En faisant le changement de variables $(x_1,\ldots,x_m)\mapsto(x_1,\ldots,x_p,y_1,\ldots,y_n)$, où $y_1 = x_1+\ldots+x_{p+1}-t$ et $y_i = x_{p+i}$ pour $i\geq 2$ (le jacobien vaut 1), on obtient

$$\alpha \;=\; \lambda^p\; e^{-\lambda t}\int_B dx_1\ldots dx_p\int_{I\!R^n_+} f(y_1,\ldots,y_n)e^{-\lambda(y_1+\ldots+y_n)}dy_1\ldots dy_n.$$

Compte tenu de (5) et de (6), on a donc $\alpha = \beta\gamma$, qui est le résultat cherché. $\square$

Ce résultat donne la "loi du processus (N_t)" au sens suivant: pour tout choix d'instants $0 < t_1 < \ldots < t_n$, on connait la loi du vecteur aléatoire $(N_{t_1},\ldots,N_{t_n})$. En effet, si $q_i\in I\!N$, on a alors

$$P(N_{t_1}=q_1,\ldots,N_{t_n}=q_n) \;=\; P(N_{t_1}=q_1,N_{t_2}-N_{t_1}=q_2-q_1,\ldots,N_{t_n}-N_{t_{n-1}}=q_n-q_{n-1}),$$

et en utilisant l'indépendance des accroissements de N_t et la proposition 2, on obtient, avec les conventions $t_0 = 0$ et $q_0 = 0$:

$$P(N_{t_1} = q_1, \ldots, N_{t_n} = q_n) \;=\; \begin{cases} \prod_{i=1}^n e^{-\lambda(t_i - t_{i-1})} \frac{(\lambda(t_i - t_{i-1}))^{q_i - q_{i-1}}}{(q_i - q_{i-1})!} & \text{si } \; q_1 \leq \ldots \leq q_n \\ 0 & \text{sinon} \end{cases} \tag{7}$$

Cette proposition admet une réciproque:

Proposition 4: *Si le processus (N_t) est à accroissements indépendants et stationnaires (au sens de la proposition 3), c'est un processus de Poisson.*

Preuve. a) Posons $S_n = T_n - T_{n-1}$ pour $n \geq 1$. Il nous faut montrer que les S_n sont indépendantes, de même loi exponentielle de paramètre λ pour un $\lambda > 0$. Pour cela, il suffit de montrer que pour chaque entier n, la variable aléatoire S_{n+1} est indépendante de $(T_1, \ldots, T_n)$ et de loi $\exp(\lambda)$.

b) Soit n fixé, et $N'_t = N_{T_n+t} - N_{T_n}$. Nous allons montrer que $(T_1, \ldots, T_n)$ et $(N'_t)_{t\geq 0}$ sont indépendants, et que la famille $(N'_t)_{t\geq 0}$ a même loi que la famille $(N_t)_{t\geq 0}$. Cela revient à montrer que

$$E(XY') \;=\; E(X)E(Y), \tag{8}$$

où $X = f(T_1, \ldots, T_n)$ avec f continue bornée sur $\mathbb{R}^n$, et $Y = g(N_{t_1}, \ldots, N_{t_q})$ et $Y' = g(N'_{t_1}, \ldots, N'_{t_q})$ pour une fonction bornée g quelconque sur $\mathbb{N}^q$ et des réels quelconques $0 < t_1 < \ldots < t_q$.

Pour tout $p \in \mathbb{N}^*$ et tout $i \in \mathbb{N}$, on pose $\alpha(i,p) = i2^{-p}$. Puis, on définit des variables aléatoires R_p en posant $R_p = \alpha(i+1,p)$ sur l'ensemble $B(p,i) = \{\alpha(i,p) \leq T_n < \alpha(i+1,p)\}$ (noter que pour chaque p, les $B(p,i)$ forment une partition de l'espace d'états Ω lorsque i décrit $\mathbb{N}$). Enfin, on pose $U_{p,j} = N_{R_p+t_j} - N_{R_p}$ et $Y'_p = g(U_{p,1}, \ldots, U_{p,q})$. Lorsque $p \uparrow \infty$ on a clairement $R_p \downarrow T$, donc à cause de la continuité à droite de $t \mapsto N_t$ on a $Y'_p \to Y'$, et les $Y'_p(\omega)$ restent bornés par une constante (indépendantes de p et de ω). La variable aléatoire X est également bornée. Par suite le théorème 4-2 implique que $E(XY'_p) \to E(XY')$. Pour obtenir (8) il suffit donc de montrer que

$$E(XY'_p) \;=\; E(X)E(Y). \tag{9}$$

Dans la suite on fixe p, et pour chaque i on pose $N(i)_t = N_{\alpha(i+1,p)+t} - N_{\alpha(i+1,p)}$. Par hypothèse les $(N(i)_t)_{t\geq 0}$ sont indépendants de la famille $\mathcal{B}p, i = (N_t)_{0\leq t\leq \alpha(i+1,p)}$ de variables aléatoires et de même loi que $(N_t)_{t\geq 0}$ (pour chaque i). Par ailleurs, sur $B_{p,i}$ on a $Y'_p = g(N(i)_{t_1}, \ldots, N(i)_{t_q})$. Enfin l'ensemble $B_{p,i}$ et la variable aléatoire X ne dépendent que de la famille $\mathcal{B}_{p,i}$ de variables aléatoires. Ainsi, (9) découle des égalités:

$$\begin{aligned} E(XY'_p) \;&=\; \sum_{i=0}^{\infty} E[Xg(N(i)_{t_1}, \ldots, N(i)_{t_q}) 1_{B_{p,i}}] \\ &=\; \sum_{i=0}^{\infty} E(X 1_{B_{p,i}}) \; E(Y) \;=\; E(X)E(Y). \end{aligned}$$

c) Avec les notations de (b), on a $S_{n+1} = \inf(t : N'_t = 1)$, donc S_{n+1} ne dépend que des $(N'_t)_{t\geq 0}$ et est défini de la même manière que S_1 à partir des $(N_t)_{t\geq 0}$. D'après (b), on a donc l'indépendance de S_{n+1} et de $(T_1, \ldots, T_n)$, et l'identité des lois de S_{n+1} et de S_1.

d) Il reste à montrer que S_1 suit une loi exponentielle. On a $\{S_1 > t\} = \{N_t = 0\}$, donc pour $s, t > 0$:

$$\begin{aligned} P(S_1 > t+s) &= P(N_{t+s} = 0) = P(N_t = 0, N_{t+s} - N_t = 0) \\ &= P(N_t = 0)P(N_s = 0) = P(S_1 > t)P(S_1 > s), \end{aligned}$$

puisque $N_{t+s} - N_t$ est indépendant de N_t et de même loi que N_s. Comme dans la preuve de la proposition 3-14, on en déduit que $P(S_1 > t) = e^{-\lambda t}$ pour un $\lambda > 0$, d'où le résultat. $\square$

4 La répartition des points d'un processus de Poisson

Pour un processus de Poisson il existe une autre description de la loi des points T_n, qui montre bien son caractère "totalement aléatoire", et que nous expliquons ci-dessous.

Soit une suite finie $V_1, \ldots, V_n$ de variables aléatoires indépendantes, uniformément distribuées sur $[0, t]$. Le vecteur $V = (V_1, \ldots, V_n)$ admet la densité $t^{-n} 1_{[0,t]^n}(x)$ sur $\mathbb{R}^n$ (cf. la proposition 3-24). On en déduit que $P(V_i = V_j) = 0$ pour tout $i \neq j$. Quitte à supprimer un ensemble de probabilité nulle, on peut donc considérer le **réarrangement croissant** $U_1, \ldots, U_n$ des V_i, c'est-à-dire la suite strictement croissante de variables aléatoires $U_1 < \ldots < U_n$ telle que les ensembles $\{U_1, \ldots, U_n\}$ et $\{V_1, \ldots, V_n\}$ soient égaux.

Pour des raisons de symétrie évidentes, on a $P(V_{\sigma_1} < \ldots < V_{\sigma_n}) = 1/n!$ pour toute permutation σ_i des n premiers entiers $\{1, \ldots, n\}$. On voit donc que le vecteur aléatoire $U = (U_1, \ldots, U_n)$ admet la densité suivante sur $\mathbb{R}^n$ (rappelons que $A_{n,t}$ est défini par (3)):

$$h_n(x) = \frac{n!}{t^n} 1_{A_{n,t}}(x). \tag{10}$$

Proposition 5: *Soit $(N_t)_{t \geq 0}$ un processus de Poisson. Conditionnellement à l'événement $\{N_t = n\}$, le vecteur aléatoire $(T_1, \ldots, T_n)$ admet la densité (10).*

Preuve. Etant donné (10), il nous suffit de montrer que pour toute fonction bornée f sur $\mathbb{R}^n$ on a

$$E[f(T_1, \ldots, T_n) 1_{\{N_t = n\}}] = P(N_t = n) \frac{n!}{t^n} \int_{A_{n,t}} f(x) dx. \tag{11}$$

Mais $\{N_t = n\} = \{T_n \leq t < T_{n+1}\}$, donc par (4) le premier membre de (11) vaut

$$\begin{aligned} &\lambda^{n+1} \int_{A_{n+1}} e^{-\lambda x_{n+1}} f(x_1, \ldots, x_n) 1_{\{x_n \leq t < x_{n+1}\}} dx_1 \ldots dx_{n+1} \\ &\qquad = \lambda^n \int_{A_{n,t}} f(x_1, \ldots, x_n) dx_1 \ldots dx_n \int_t^\infty \lambda e^{-\lambda y} dy, \end{aligned}$$

et comme $P(N_t = n) = e^{-\lambda t} (\lambda t)^n / n!$ on a (11). $\square$

Etant donnée la proposition 3, le même résultat subsiste pour les points tombant dans n'importe quel intervalle, et de plus ce qui se passe dans des intervalles disjoints est indépendant.

Ainsi, pour un processus de Poisson, toute se passe comme si on faisait la construction suivante: on se donne une suite $(Q_p)_{p\in I\!N}$ de variables indépendantes de loi de Poisson de paramètre λ. Puis dans chaque intervalle $]p,p+1]$ on "jette" Q_p points, indépendamment les uns des autres, selon la loi uniforme sur cet intervalle. L'ensemble des T_n pour $n \geq 1$ est alors la réunion de ces points, numérotés selon les abscisses croissantes. Inversement, cette construction conduit toujours à un processus de Poisson, comme le montre la

Proposition 6: *Avec la construction précédente, si $T_0 = 0$ et si N_t est défini par (1), le processus $(N_t)_{t\geq 0}$ est un processus de Poisson.*

Preuve. Il suffit de montrer que (N_t) est un processus à accroissements indépendants et stationnaires. Pour cela, il suffit en fait de vérifier que le nombre de points tombés dans des intervalles deux-à-deux disjoints sont des variables aléatoires indépendantes, dont les lois ne dépendent que des longueurs des intervalles.

Ce qui se passe sur des intervalle $]p,p+1]$ différents est par construction indépendant et de même loi. Il suffit donc de montrer la propriété précédente, pour des intervalles deux-à-deux disjoints, tous contenus dans un même $]p,p+1]$, et on peut même prendre $p = 0$.

Soit donc $I_1,\ldots,I_q$ des sous-intervalles deux-à-deux disjoints de $I =]0,1]$, de longueurs respectives $a_1,\ldots,a_q$. Quitte à rajouter un intervalle supplémentaire, on peut même supposer que la réunion des I_i égale I, donc $a_1+\ldots+a_q = 1$. Soit M_i le nombre de points tombés dans I_i. On note aussi X_i^n la variable aléatoire qui vaut 1 si le $n^{\text{ème}}$ point jeté dans I tombe dans I_i, et 0 sinon. On a donc $M_i = X_i^1+\ldots+X_i^{Q_0}$ (avec $M_i = 0$ si $Q_0 = 0$). Les variables aléatoires $(X_1^n,\ldots,X_q^n)_{n\geq 1}$ sont indépendantes, ce qui entraine (exactement comme pour (6-12)) que si $Q_0 = n$, la variable aléatoire $(M_1,\ldots,M_n)$ suit une loi multinomiale de taille n et de paramètres les a_i. Il vient alors pour tous entiers n_i, et avec $n = n_1+\ldots+n_q$:

$$\begin{aligned} P(M_1 = n_1,\ldots,M_q = n_q) &= P(Q_0 = n)P(M_1 = n_1,\ldots,M_q = n_q/Q_0 = n) \\ &= e^{-\lambda}\,\frac{\lambda^n}{n!}\,\frac{n!}{n_1!\ldots n_q!}\,a_1^{n_1}\ldots a_q^{n_q} \\ &= \prod_{i=1}^{q} e^{-\lambda a_i}\,\frac{(\lambda a_i)^{n_i}}{n_i!}. \end{aligned} \tag{12}$$

On en déduit que les M_i sont indépendants et que chaque M_i suit la loi de Poisson de paramètre λa_i: on a donc le résultat. □

5 Le "paradoxe de l'autobus"

Nous terminons ce chapitre par un résultat curieux, appelé le "paradoxe de l'autobus": on suppose que les points T_n d'un processus de Poisson (pour $n \geq 1$) représentent les instants d'arrivée des autobus à un arrêt, bien que comme on l'a déjà dit cette modélisation n'est pas réellement adéquate pour représenter ce phénomène aléatoire...

On considère un voyageur qui arrive à l'arrêt à l'instant t. On note U_t son temps d'attente, c'est-à-dire $U_t = \inf(T_n : T_n > t, n \geq 1) - t$. On note aussi $V_t = t - \sup(T_n :$

$T_n \leq t, n \geq 0$) le temps qui sépare la dernière arrivée d'autobus avant t, et t (avec la convention $V_t = t$ s'il n'est pas arrivé d'autobus avant t). Soit enfin $W_t = U_t + V_t$ le temps entre les deux autobus arrivant juste avant et juste après t.

Proposition 7: *Les variables aléatoires U_t et V_t sont indépendantes. La variable U_t suit une loi exponentielle de paramètre λ, et V_t suit une loi exponentielle "tronquée" de fonction de répartition*

$$F(x) = \begin{cases} 0 & si \;\; x < 0 \\ 1 - e^{-\lambda x} & si \;\; 0 \leq x < t \\ 1 & si \;\; x \geq t. \end{cases}$$

Preuve. On a $V_t \leq t$ par construction, et si $y \geq 0$ et $x \in [0, t[$ on a $U_t > y$ et $V_t \geq x$ si et seulement si $N_{t+y} - N_{t-x} = 0$. Comme cette dernière variable aléatoire suit une loi de Poisson de paramètre $\lambda(x+y)$, on en déduit que

$$P(U_t > y, V_t \geq x) = e^{-\lambda(x+y)}.$$

Le résultat en découle très simplement. $\square$

Ce résultat est considéré comme un paradoxe pour la raison suivante: d'une part, les intervalles de temps séparant deux arrivées successives d'autobus sont exponentiels de paramètre λ, et donc de moyenne $1/\lambda$. Le temps d'attente d'un client arrivant en t suit la même loi, alors qu'on s'attendrait *a priori* à ce qu'il attende "en moyenne" $1/2\lambda$. Et, lorsque t est grand, V_t est aussi approximativement exponentielle de paramètre λ (i.e., lorsque $t \to \infty$, les variables aléatoires V_t convergent en loi vers une variable aléatoire exponentielle de paramètre λ): donc pour t "grand", l'intervalle $W_t = U_t + V_t$ suit une loi $\Gamma(2, \lambda)$, de moyenne $2/\lambda$. Donc, bien que l'intervalle moyen entre deux arrivées soit $1/\lambda$, l'intervalle moyen entre deux arrivées pour un usager donné est $2/\lambda$! cela s'explique par le fait qu'un usager a "plus de chances" d'arriver à l'intérieur d'un intervalle de grande taille qu'à l'intérieur d'un intervalle de petite taille.