Bruits et Signaux.

Didier Pelat

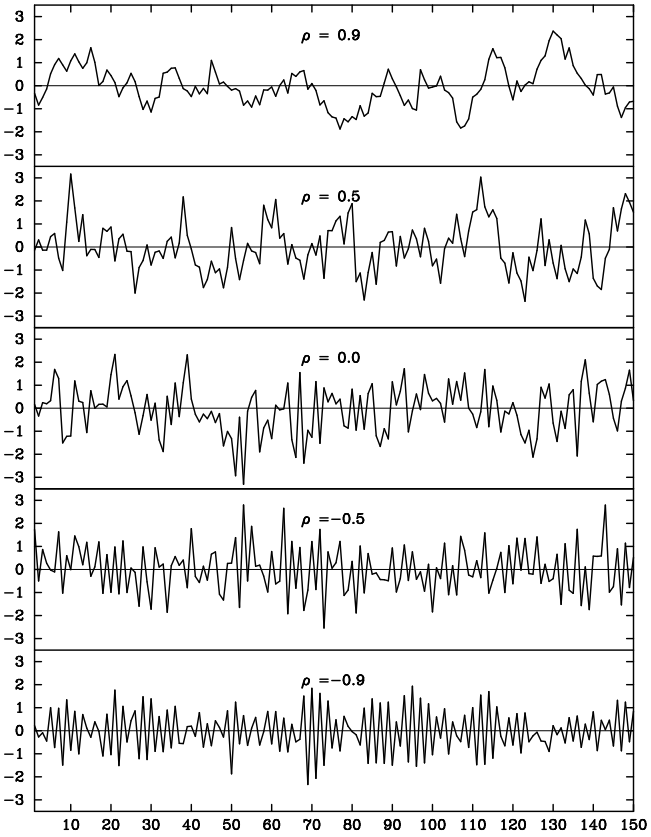
Ecole doctorale d'Ile de France.

D.E.A. d'Astrophysique et Techniques Spatiales,
D.E.A. d'Astronomie et Techniques Spatiales.

Cours

BRUITS et SIGNAUX

(Introduction aux méthodes de traitements des données.)



par
Didier PELAT

Note à l'intention des lecteurs.

Un cours intitulé « Bruits et Signaux », devrait idéalement comporter au moins deux parties, une partie traitant des statistiques des variables aléatoires, et une autre partie traitant des statistiques des processus stochastiques. On ne trouvera ici que la partie concernant les variables aléatoires, mais les principes de base exposés ici seront un bagage précieux pour quiconque voudra bien aborder des problèmes plus complexes.

Quelques chapitres contiennent des programmes écrit en FORTRAN dont le but est d'illustrer tel ou tel point qui vient d'être traité. L'auteur a la conviction que de tels programmes possèdent une valeur pédagogique mais il est également persuadé qu'ils sont imparfaits, en conséquence il décline toute responsabilité en cas d'usage de ces programmes.

C'est un plaisir de remercier ici D. Alloin, J. Ballet, L. Carter et les étudiants de l'école doctorale d'île-de-France pour leurs lectures critiques de ce texte. Je tiens tout particulièrement à remercier Françoise Launay pour avoir relu entièrement ce texte et y avoir apporté d'innombrables corrections.

Tous commentaires, corrections de fautes de frappe etc... sont vivement encouragés.

Meudon le 15 octobre 1998

D. PELAT
D.A.E.C.
Observatoire de PARIS section de MEUDON
92195–MEUDON CEDEX

Tel : 01 45 07 74 37
Fax : 01 45 07 74 69
Internet : pelat@mesioc.obspm.fr

Table des matières

Table des figures

Liste des tableaux

Première partie

Eléments de théorie des probabilités.

Chapitre 1

Espaces probabilisés.

La base fondamentale du calcul des probabilités est la théorie de la mesure élaborée par Borel et Lebesgue au début de ce siècle. On doit à Kolmogorov, dans les années 1930, d'avoir reconnu qu'une probabilité se concevait en tant que « mesure » d'une certaine classe d'événements. Avant cette date la théorie des probabilités n'avait pas le statut de théorie mathématique cohérente et certaines notions comme celle de probabilité conditionnelle restaient assez vagues.

1.1 Les axiomes de Kolmogorov.

Suivant Kolmogorov, un espace probabilisé est un triplet $(\Omega, \mathfrak{B}, \mathrm{Pr})$ constitué : 1) d'un ensemble Ω dont les éléments ω sont appelés *événements élémentaires* ; 2) d'une classe $\mathfrak{B}$ possédant une structure dite de *tribu* dont les éléments sont appelés *événements* ; et 3) d'une mesure normée Pr dite de *probabilité* sur les éléments de cette tribu.

On rappelle qu'une classe est un ensemble de parties (c'est-à-dire de sous-ensembles) d'un ensemble.

1.1.1 L'ensemble Ω.

L'ensemble Ω est dit *ensemble des épreuves*. Ses éléments ω sont toutes les issues possibles d'une expérience soumise au hasard. Ces issues ou *événements élémentaires* ou encore *événements atomiques* sont de nature abstraite. Ce sont par exemple, les côtés d'une pièce de monnaie dans le jeu de pile ou face, Ω = {pile, face}, ou les caractères génétiques d'un individu, Ω ={yeux bleus, yeux verts,... cheveux blonds, cheveux roux,...}. Un grand nombre d'expériences suscitent des événements élémentaires de nature éminemment numérique, comme le jet d'un dé à 6 faces, $\Omega = \{1, 2, 3, 4, 5, 6\}$ ou la mesure de la taille d'un individu, dont le résultat est un réel strictement positif, $\Omega = \mathbb{R}^+$.

Les notions introduites par la théorie des ensembles s'appliquent naturellement à l'ensemble Ω, à ses éléments et à ses parties, mais la théorie des probabilités utilise une terminologie particulière, résumée par le tableau 1.1.

Notation	Terminologie ensembliste	Terminologie probabiliste
$\omega \in \Omega$	élément	événement élémentaire
$A \subseteq \Omega$	partie, sous-ensemble	événement
Ω	partie pleine	événement certain
$\varnothing$	partie vide	événement impossible
$A \cap B$ ou AB	intersection de A et B	A et B sont simultanés
$A \cup B$	réunion de A et B	A et/ou B est réalisé
$\Omega - A$ ou A^c	complémentaire de A	événement contraire de A
$A \subseteq B$ ou $A \Rightarrow B$	A est inclus dans B	A implique B
$AB = \varnothing$	A et B sont disjoints	A et B sont incompatibles

TAB. 1.1: *Tableau comparatif des terminologies ensembliste et probabiliste, inspiré de Charles B. et Escoufier Y., (1971) [16].*

1.1.2 La tribu $\mathfrak{B}$.

Une tribu est un ensemble dont les éléments sont des parties de Ω. Cet ensemble sera appelé *tribu* si ses éléments obéissent à un certain nombre de règles définies ci-après. Les parties de Ω appartenant à une tribu seront appelés *événements*. On notera habituellement les événements de Ω par des capitales romaines : $A, B, \ldots$ et par le symbole $\{A_k\}, (k = 1, \ldots, \infty)$, une suite d'événements de Ω.

Opérations sur les événements. On suppose que l'on a défini entre événements deux opérations commutatives et associatives notées $\cup$ et $\cap$, distributives l'une par rapport à l'autre :

$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C), \quad A \cap (B \cup C) = (A \cap B) \cup (A \cap C)\,. \tag{1.1}$$

De plus il existe deux événements particuliers appelés événement *certain* et événement *impossible*, notés respectivement Ω et $\varnothing$. Ce sont les solutions uniques des équations suivantes :

$$A \cup \varnothing = A, \quad A \cap \varnothing = \varnothing, \quad A \cup \Omega = \Omega, \quad A \cap \Omega = A\,. \tag{1.2}$$

Enfin il existe une opération de *complémentarité* qui à un événement A lui associe l'événement A^c dit *événement contraire*. Cet événement est l'unique solution des équations suivantes :

$$A \cup A^c = \Omega, \quad A \cap A^c = \varnothing, \quad (A^c)^c = A\,. \tag{1.3}$$

Il résulte de ces définitions que $\Omega^c = \varnothing$ et $\varnothing^c = \Omega$, et que les opérations $\cup$ et $\cap$ satisfont la règle de « de Morgan » :

$$(A \cup B)^c = A^c \cap B^c, \quad (A \cap B)^c = A^c \cup B^c\,. \tag{1.4}$$

Le sens qu'il convient de donner à ces opérations est le suivant : on dit que l'événement $A \cup B$ est réalisé si l'issue ω de l'épreuve appartient à A ou à B

$$\omega \in A \cup B \iff \omega \in A \text{ ou } \omega \in B\,, \tag{1.5}$$

on dit que l'événement $A \cap B$ est réalisé si ω appartient à A *et* à B,

$$\omega \in A \cap B \iff \omega \in A \text{ et } \omega \in B\,. \tag{1.6}$$

On note habituellement AB l'événement $A \cap B$ et on dit que les événements A et B sont simultanés. Si $AB = \varnothing$, on dit que A et B sont des événements *incompatibles*, dans ce cas leur union est souvent notée $A + B$.

Remarque 1.1. La démarche qui consiste à admettre que si l'événement A manque de se réaliser, c'est que sont contraire A^c s'est réalisé, ne s'est imposé aux esprits qu'au XVIII^e siècle. Il semble que ce soit Th. Bayes qui, le premier, ait admis l'équivalence entre l'échec d'un événement et le succès de son contraire (voir, A.I. Dale § 2.3 [18]).

Système complet d'événements. Une suite d'événements $\{A_k\}$ forme un *système complet d'événements* si les éléments qui la composent sont non-vides, deux à deux incompatibles, et si leur union (éventuellement infinie) recouvre tout Ω, c'est-à-dire si :

$$\forall i, j = 1, \ldots, \infty,\ i \neq j; \quad A_i A_j = \varnothing \quad \text{et} \quad \bigcup_{k=1}^{\infty} A_k = \Omega\,. \tag{1.7}$$

Ainsi tous les événements élémentaires ω de Ω appartiennent à un et un seul événement d'un système complet.

▶**Exemple 1.1.** *Systèmes complets sur des espaces des épreuves finis.* Si Ω n'est formé que du seul événement élémentaire ω_1, il n'y a qu'un seul système complet : $\{\{\omega_1\}\}$; si $n = 2$, il y a deux systèmes complets : $\{\{\omega_1, \omega_2\}\}$ et $\{\{\omega_1\}, \{\omega_2\}\}$; si $n = 3$ il y en a 5 : $\{\{\omega_1, \omega_2, \omega_3\}\}$, $\{\{\omega_1, \omega_2\}, \{\omega_3\}\}$, $\{\{\omega_1, \omega_3\}, \{\omega_2\}\}$, $\{\{\omega_2, \omega_3\}, \{\omega_1\}\}$ et enfin $\{\{\omega_1\}, \{\omega_2\}, \{\omega_3\}\}$. Voir exercice 1.2

Algèbre d'événements. Pour qu'un ensemble d'événements de Ω soit une tribu, il faut d'abord qu'il soit une *algèbre de Boole*. Un ensemble d'événements $\mathcal{A}$ constitue une algèbre de Boole, ou plus simplement une algèbre, si les conditions suivantes sont satisfaites :

A1 : $\Omega \in \mathcal{A}$

A2 : $A \in \mathcal{A} \Rightarrow A^c \in \mathcal{A}$ donc $\varnothing = \Omega^c \in \mathcal{A}$

A3 : $A_1, A_2 \in \mathcal{A} \Rightarrow A_1 \cup A_2 \in \mathcal{A}$ donc $A_1 A_2 = (A_1^c \cup A_2^c)^c \in \mathcal{A}$

▶**Exemple 1.2.** Quelque soit Ω, on a toujours les deux algèbres suivantes :

$$P(\Omega), \quad \textit{l'ensemble des parties de } \Omega,$$
$$G(\Omega) = \{\varnothing, \Omega\},$$

$P(\Omega)$ est dite algèbre *discrète* et $G(\Omega)$ algèbre *grossière*.

Pour les ensembles Ω formés d'un nombre fini d'éléments, il n'y a aucune difficulté à considérer n'importe quelle partie de Ω comme constituant un événement. En revanche, pour des ensembles Ω infinis, il n'est pas possible d'affecter une mesure cohérente à toute partie de Ω (voir remarque 1.2). Il faut donc définir une structure susceptible de comprendre la classe des parties de Ω (pour le cas fini), mais pas uniquement celle-ci. On doit, d'un côté, restreindre les

parties de Ω qui peuvent appartenir à la structure à celles qui conduisent à une mesure cohérente ; mais, d'un autre côté, cette structure doit être suffisamment riche de façon à inclure des événements aussi intéressants que ceux portant sur des limites de suites infinies d'événements (voir exemple 1.5). Ce double objectif a été atteint par Kolmogorov qui a introduit la notion de tribu dans un ouvrage paru en 1933 [43].

Tribu d'événements. L'algèbre $\mathcal{A}$ devient une tribu $\mathfrak{B}$ (ou σ-algèbre) si l'union d'une infinité *dénombrable* d'événements est aussi un événement. C'est-à-dire :

B3 : $A_k \in \mathfrak{B}, (k = 1, \ldots, \infty) \Rightarrow \bigcup_{k=1}^{\infty} A_k \in \mathfrak{B}$

Il résulte des axiomes que l'intersection d'une suite infinie d'événements appartenant à une tribu, appartient aussi à la tribu :

$$A_k \in \mathfrak{B},\ (k = 1, \ldots, \infty) \Rightarrow \bigcap_{k=1}^{\infty} A_k \in \mathfrak{B}\,. \tag{1.8}$$

On remplace souvent l'axiome **A2** par la formule (1.8), et les axiomes définissant une tribu d'événements deviennent alors :

B1 : $\Omega \in \mathfrak{B}$,

B2 : $A_k \in \mathfrak{B}, (k = 1, \ldots, \infty) \Rightarrow \bigcap_{k=1}^{\infty} A_k \in \mathfrak{B}$,

B3 : $A_k \in \mathfrak{B}, (k = 1, \ldots, \infty) \Rightarrow \bigcup_{k=1}^{\infty} A_k \in \mathfrak{B}$.

Remarque 1.2. Cette façon de définir une tribu permet de faire un parallèle intéressant entre une tribu d'événements et une topologie d'ouverts (voir § B.2.1, page 317).

On notera, en particulier, que l'on autorise les réunions infinies *non dénombrables* d'ouverts alors que l'on n'autorise que les réunions d'une infinité *dénombrale* d'événements. Une telle restriction vient de la théorie de la mesure et répond au souci de donner une mesure cohérente aux parties de Ω entrant dans la tribu. Si $\Omega = \mathbb{R}^3$, on peut donner une mesure à une surface, en cm^2 par exemple, mais un volume est une réunion continue de surfaces, et admettre cette réunion dans la tribu conduirait à mesurer des volumes en cm^2, ce qui est physiquement incohérent. L'appartenance à une tribu assure à tous ses membres la possibilité d'avoir une mesure homogène.

▶**Exemple 1.3.** Les classes de parties suivantes sont des tribus de Ω :

1. $P(\Omega)$, l'ensemble des parties de Ω. C'est utile pour le cas fini, cela l'est moins pour le cas infini.
2. $G(\Omega) = \{\varnothing, \Omega\}$, la tribu grossière.
3. Si les événements A_j forment un système complet *dénombrable*, la classe des unions des A_j forment une tribu. Par exemple tous les événements suivant appartiennent à cette tribu :

$$A_1,\ A_2,\ A_3, \ldots$$
$$A_1 \cup A_2,\ A_1 \cup A_3,\ A_1 \cup A_4,\ A_2 \cup A_3, \ldots$$
$$A_1 \cup A_2 \cup A_3,\ A_1 \cup A_2 \cup A_4,\ A_1 \cup A_2 \cup A_5,\ A_1 \cup A_3 \cup A_4, \ldots$$

4. On montre que pour toute classe de parties de Ω, il existe une plus petite tribu (au sens de l'inclusion) qui contienne cette classe. On l'appelle *tribu engendrée* par la classe.

1.1.3 La mesure de probabilité Pr.

Un espace tel que Ω pour lequel on a défini une tribu est appelé espace *mesurable*, pour que Ω devienne un espace *probabilisé* il faut définir de plus une mesure particulière appelée *probabilité*. Une probabilité Pr est une application qui associe un nombre à un événement d'une tribu $\mathfrak{B}$ et qui possède les propriétés **P1**, **P2** et **P3** suivantes :

P1 : $A \in \mathfrak{B} \Rightarrow \Pr\{A\} \geq 0$ (*positivité.*)

La probabilité doit être additive pour des événements incompatibles, c'est-à-dire : $AB = \varnothing \Rightarrow \Pr\{A \cup B\} = \Pr\{A\} + \Pr\{B\}$. Il faut de plus qu'elle soit σ-additive, c'est-à-dire que si la suite dénombrable $A_k, (k = 1, \ldots, \infty)$ est composé d'événements disjoints, alors :

P2 : $\Pr\Big\{\bigcup_{k=1}^{\infty} A_k\Big\} = \sum_{k=1}^{\infty} \Pr\{A_k\}$ (*additivité dénombrable.*)

Pour que la mesure Pr soit une probabilité il faut, finalement, qu'elle soit normalisée :

P3 : $\Pr\{\Omega\} = 1$ (*normalisation.*)

1.1.4 Exemples

Lorsqu'il s'agit d'utiliser dans le monde réel le modèle théorique exposé ci-dessus, il se pose alors le problème du choix de chacun des termes du triplet $(\Omega, \mathfrak{B}, \Pr)$. Ce choix n'est pas imposé par la théorie des probabilités, il est le résultat d'une analyse critique du phénomène que l'on cherche à modéliser. De vigoureux débats ont eu lieu par le passé entre les avocats de tel ou tel modèle. Dans une certaine mesure ce débat dure encore aujourd'hui, et porte sur le problème de l'interprétation pratique de la notion de probabilité (voir remarque 1.3, page 11).

▶**Exemple 1.4.** *Espace fini.* Si Ω contient un nombre fini d'éléments, la tribu $\mathfrak{B}$ est presque toujours formée des parties de Ω et la mesure de probabilité Pr est souvent constante. La plupart du temps on a d'ailleurs choisi Ω pour qu'il en soit ainsi.

Dans l'expérience du jet de 2 dés à 6 faces, on peut choisir les modèles suivants.

1. Ω est formé de tous les couples ordonnés formés par les chiffres portés par les dés : $\Omega = \{1, 2, 3, 4, 5, 6\}^2$. La tribu est formée des parties de Ω, par exemple $A = \{(1, 3), (3, 1), (2, 2)\}$ = « la somme vaut 4 ». La mesure de probabilité est égale à $\frac{1}{36}$, pour tous les éléments de Ω. Dans ce modèle, les dés sont *discernables*.
2. On peut accepter une description moins fine du phénomène en ne considérant que la somme des chiffres portés sur les dés. On a $\Omega = \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$ et $\Pr = \{\frac{1}{36}, \frac{2}{36}, \frac{3}{36}, \frac{4}{36}, \frac{5}{36}, \frac{6}{36}, \frac{5}{36}, \frac{4}{36}, \frac{3}{36}, \frac{2}{36}, \frac{1}{36}\}$.
3. On peut aussi choisir le modèle, $\Omega = \{1, 2, 3, 4, 5, 6, \text{« cassé »}\}^2$ pour lequel il n'y a pas de mesure de probabilité intuitive.

4. Un choix moins heureux aurait été celui défendu par le « *Chevalier de Méré* »[1] où les dés sont *indiscernables*. On aurait $\Omega = \{2,3,4,5,6,7,\ 8,9,10,11,12\}$, et $\Pr = \{\frac{1}{21}, \frac{1}{21}, \frac{2}{21}, \frac{2}{21}, \frac{3}{21}, \frac{3}{21}, \frac{3}{21}, \frac{2}{21}, \frac{2}{21}, \frac{1}{21}, \frac{1}{21}\}$.

Le modèle 4 ne correspond pas à l'idée que l'on se fait de l'expérience en question[2] et il n'est pas confirmé par la pratique. Pour le modèle 3, on obtient un modèle utilisable en rejetant l'événement « cassé » hors de Ω. Le modèle 2 est après tout celui dont on doit souvent se contenter dans le monde réel : on ne connaît pas toujours tous les paramètres d'une expérience. Le modèle 1 est le *bon* modèle mais il est rare que l'on ait affaire à un cas si pur.

▶**Exemple 1.5.** *Espace infini.* Considérons le jeu de « pile ou face » infini. Un événement élémentaire ω est constitué d'une suite infinie de « pile » et de « face ». Affectons à « pile » la valeur 1 et à « face » la valeur 0. Posons $X_n(\omega)$ égal à l'issue du n^{e} tirage, $X_n = 1$ si on a obtenu « pile » au n^{e} tirage, $X_n = 0$ si c'est « face ».

On choisit pour espace des épreuves, toutes les suites infinies de 0 et de 1 : $\Omega = \{0,1\}^\infty$. On peut alors considérer un événement élémentaire comme la partie fractionnaire écrite en binaire d'un nombre réel x appartenant à l'intervalle $[0,1[$. On peut choisir pour algèbre des d'événements les familles de parties de Ω suivantes :

1. Les intervalles quelconques (ouverts, fermés et semi-ouverts) de $[0,1]$. Ce choix peut être utile si le but du tirage aléatoire est de déterminer « au hasard » un nombre x compris entre 0 et 1.
2. Si l'on s'intéresse plutôt à des questions comme « quelle est la probabilité en fonction de n d'obtenir k piles consécutifs au cours de n tirages ? » on aura intérêt à choisir $\forall n$, la famille des parties de $\{0,1\}^n$. Par exemple, pour $n = 1$ et $n = 2$, on a les familles de parties $P_1(\Omega)$ et $P_2(\Omega)$:

$$P_1(\Omega) = \{\varnothing, \{0\}, \{1\}, \{0,1\}\},$$
$$P_2(\Omega) = \{(\varnothing,\varnothing), (\varnothing,\{0\}), (\varnothing,\{1\}), (\varnothing,\{0,1\}), (\{0\},\varnothing), \ldots, (\{0,1\},\{0,1\})\}.$$

Il est clair que ces familles constituent des algèbres, mais il existe des événements « intéressants » qui ne sont pas dans ces algèbres. Par exemple l'événement L, qui concerne ce que l'on appelle la *loi des grands nombres* :

$$L = \{\omega \mid \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} X_i(\omega) = \tfrac{1}{2}\}.$$

Un événement élémentaire ω appartient à L si le nombre moyen de « pile » est égal au nombre moyen de « face » lorsque le nombre d'épreuves croît au delà de toute limite. Par exemple, si $x \in [0,1[$ représente l'issue d'une épreuve, les deux rationnels $x = \frac{1}{3}$ et $x = \frac{2}{3}$ appartiennent à L, il semble bien que π appartienne aussi à L, en revanche $x = \frac{1}{7}$ n'y appartient pas. Quoi qu'il en soit[3], L n'appartient pas à l'une ou l'autre des deux algèbres définies ci-dessus mais, L est exprimable à l'aide d'unions et d'intersections dénombrables d'éléments de ces algèbres. C'est donc un événement. En effet, on a :

$$L = \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} \{\omega \mid \frac{1}{2k} \sum_{i=1}^{2k} X_i(\omega) = \tfrac{1}{2}\},$$

La probabilité de L vaut soit 0 soit 1, elle ne vaut 1 que si le jeu est *équitable*, c'est-à-dire si : $\Pr\{\text{« pile »}\} = \Pr\{\text{« face »}\} = \frac{1}{2}$.

1. Voir la lettre de Pascal à Fermat du 29 juillet 1654 [56].

2. Ce n'était pas le cas de Leibnitz qui, comme Méré, pensait que 11 était aussi probable que 12 car il ne distinguait pas l'issue (5,6) de l'issue (6,5).

3. Borel a montré en 1909 que presque tous les nombres réels étaient « normaux », c'est-à-dire que les proportions des chiffres formant ces nombres, exprimés en quelque base que ce soit, étaient égales. Pour nous cela veut dire que presque tout les x appartiennent à L

1.1.5 Ensemble de mesure nulle.

Etant donné un espace probabilisé $(\Omega, \mathfrak{B}, \Pr)$, on dit que l'événement $A \in \mathfrak{B}$ est de mesure nulle si $\Pr\{A\} = 0$. On dira que l'événement A est *presque sûr* si $\Pr\{A^c\} = 0$, c'est-à-dire si l'événement non-A possède une mesure nulle. De même on dira qu'une propriété a lieu *presque partout* relativement à Pr si le sous-ensemble A des ω qui vérifient cette propriété est un événement presque-sûr. Autrement dit, une propriété a lieu presque-partout si sa non-satisfaction possède une probabilité nulle de se réaliser.

1.2 Probabilités conditionnelles.

Lorsque l'on ne possède pas d'information sur un événement A, autre que c'est bien un événement ($A \in \mathfrak{B}$), les probabilités qui satisfont les axiomes ci-dessus sont dites probabilités *a priori*. Si l'on a connaissance de la probabilité d'un événement B, il est possible de définir une autre mesure $\Pr_B$, définie par rapport à $\Pr\{B\}$:

$$\Pr_B\{A\} = \frac{\Pr\{AB\}}{\Pr\{B\}} . \tag{1.9}$$

Si $\Pr\{B\} \neq 0$, il est facile de vérifier que cette nouvelle mesure satisfait les axiomes **P1**, **P2**, **P3** et que par conséquent c'est une probabilité. De telles probabilités sont appelées probabilités *a posteriori* ou probabilités *conditionnelles* (conditionnellement à B). On note plus couramment $\Pr\{A|B\}$ cette probabilité, ce qui autorise l'écriture :

$$\Pr\{AB\} = \Pr\{B\} \Pr\{A|B\} . \tag{1.10}$$

$\Pr\{A|B\}$ se dit probabilité de A *sachant* B. Cette mesure n'est autre que la mesure normalisée de la partie de A qui est incluse dans B. La probabilité *a priori* $\Pr\{A\}$ est en général différente de la probabilité conditionnelle $\Pr\{A|B\}$, cette différence révèle ce que l'on appelle un *effet de sélection* ou *conditionnement*.

L'interprétation pratique des probabilités conditionnelles est particulièrement simple lorsque l'espace des épreuves Ω est fini et lorsque des considérations de symétrie permettent d'affecter une probabilité égale à tous les événements élémentaires ω de Ω. Dans ce cas, la probabilité conditionnelle n'est autre que la proportion des ω de A qui sont dans B.

Formule de Bayes.

De la même façon que l'on a défini une probabilité conditionnelle $\Pr\{A|B\}$, on peut, sous réserve que $\Pr\{A\} \neq 0$, définir $\Pr\{B|A\}$. Il vient alors : $\Pr\{AB\} = \Pr\{B\} \Pr\{A|B\} = \Pr\{A\} \Pr\{B|A\}$ d'où la formule dite de Bayes :

$$\Pr\{B|A\} = \frac{\Pr\{A|B\} \Pr\{B\}}{\Pr\{A\}} . \tag{1.11}$$

Les idées conduisant à cette formule ont été introduites pour la première fois par Th. Bayes dans un essai posthume publié en 1764 [4].

Formule de « probabilité des causes ».

La formule de Bayes est utilisée dans un contexe où A représente une donnée connue et B une conjecture sur l'état de la nature susceptible d'avoir conduit à A. Notons plutôt X l'événement qui représente l'ensemble des données connues sur « l'état de la nature », et par $H_0, H_1, \ldots, H_n$ différentes conjectures sur cet état. La formule de Bayes s'écrit alors :

$$\Pr\{H_i|X\} = \frac{\Pr\{X|H_i\}\Pr\{H_i\}}{\Pr\{X\}}\,. \tag{1.12}$$

La probabilité *a posteriori* $\Pr\{H_i|X\}$ est la probabilité pour que la nature soit dans l'état i étant donné l'information X que l'on a sur elle, $\Pr\{H_i\}$ est la probabilité *a priori* pour que la nature soit dans cet état i et $\Pr\{X|H_i\}$ est la probabilité d'obtenir X lorsque l'on suppose que la nature est dans l'état i. Cette dernière probabilité reçoit le nom de *vraisemblance* de l'hypothèse H_i vis-à-vis des données X.

Si les hypothèses $H_0, \ldots, H_n$ forment un système complet d'événements disjoints, c'est-à-dire si une et une seule d'entre elles est vraie, alors $\Pr\{X\}$ peut s'écrire : $\Pr\{X\} = \Pr\{\cup_{i=1}^n X H_i\}$ et d'après l'équation (1.10) et l'axiome **P2** il vient :

$$\Pr\{H_i|X\} = \frac{\Pr\{X|H_i\}\Pr\{H_i\}}{\sum_{i=1}^n \Pr\{X|H_i\}\Pr\{H_i\}}\,. \tag{1.13}$$

Cette formule appelée formule de la *probabilité des causes* permet de calculer les probabilités *a posteriori* connaissant les probabilités *a priori* et les vraisemblances des diverses hypothèses vis-à-vis de l'information représentée par X.

▶**Exemple 1.6.** *Modèle des urnes.* Il est traditionnel de représenter les divers « états de la nature » par des urnes contenant des boules blanches et des boules noires en diverses proportions. Considérons un problème à deux états :

1. suivant l'hypothèse H_0, l'expérimentateur a devant lui l'urne n° 0 contenant autant de boules blanches que de boules noires ;
2. suivant l'hypothèse H_1, il a devant lui l'urne n° 1 contenant trois fois plus de boules noires que de boules blanches.

On a placé une urne au hasard devant l'expérimentateur et il en a extrait une boule noire. On demande la probabilité pour que l'urne d'où a été extraite la boule soit respectivementment l'urne n° 0 ou l'urne n° 1.

La donnée X sur l'état de la nature est la boule noire extraite de l'urne, il est facile de trouver les vraisemblances des diverses hypothèses. Si l'urne choisie est l'urne n° 0, la probabilité d'en extraire une boule noire est $\frac{1}{2}$; si c'est l'urne n° 1 cette probabilité vaut $\frac{3}{4}$, on a donc :

$$\Pr\{X|H_0\} = \frac{1}{2}\,, \quad \Pr\{X|H_1\} = \frac{3}{4}\,.$$

Si les urnes ont été choisies à « pile » ou « face », par exemple, on a les probabilités *a priori* : $\Pr\{H_0\} = \Pr\{H_1\} = \frac{1}{2}$, et on en déduit les probabilités *a posteriori* demandées

$$\Pr\{H_0|X\} = \frac{\frac{1}{2}\times\frac{1}{2}}{\frac{1}{2}\times\frac{1}{2}+\frac{3}{4}\times\frac{1}{2}} = \frac{2}{5}\,, \quad \Pr\{H_1|X\} = \frac{\frac{3}{4}\times\frac{1}{2}}{\frac{1}{2}\times\frac{1}{2}+\frac{3}{4}\times\frac{1}{2}} = \frac{3}{5}\,.$$

L'information X est marginalement favorable à l'hypothèse que l'urne placée devant l'expérimentateur était l'urne n° 1. Ce résultat dépend fortement des valeurs attribuées aux probabilités *a priori*.

Remarque 1.3. Utilisation bayesienne de la formule de Bayes. La formule de Bayes et celle de la probabilité des causes sont des conséquences assez élémentaires des axiomes de définition, en cela elles ne sont pas criticables. Un problème apparaît cependant lorsqu'on veut les utiliser pour valider telle ou telle hypothèse H_i au vu des données X. Une telle opération n'est possible que si l'on connait les probabilités *a priori* $\Pr\{H_i\}$, c'est-à-dire, la probabilité pour que la nature soit dans l'état *i en l'absence de toute information* sur elle.

Les praticiens se séparent en une école *classique* qui refuse d'affecter une valeur aux probabilités *a priori* et une école *bayesienne* qui l'accepte en élargissant la notion de probabilité à celle de *plausibilité*. Selon cette école, la probabilité élargie à la plausibilité mesure un degré subjectif de croyance envers l'état de la nature avant toute expérimentation. Un tel point de vue est discutable mais il est opérationnel dans le sens où les probabilités *a priori* et l'information X permettent de calculer des probabilités *a posteriori* qui deviendront les probabilités *a priori* relativement à une nouvelle information X'. L'arbitraire initial est en quelque sorte « oublié » au fur et à mesure que l'on gagne de l'information.

1.3 Evénements indépendants.

Nous dirons que deux événements A et B sont *indépendants* si, et seulement si, on peut écrire :

$$\Pr\{AB\} = \Pr\{A\}\Pr\{B\}\,. \tag{1.14}$$

Si l'événement B n'est pas l'événement impossible, $\Pr\{B\}$ n'est pas nul et il vient d'après (1.11) : $\Pr\{A|B\} = \Pr\{A\}$; réciproquement, si $\Pr\{A\} \neq 0$, on a $\Pr\{B|A\} = \Pr\{B\}$. Ces relations sont conformes à la notion intuitive d'indépendance : si A et B sont indépendants, alors la réalisation de l'un n'affecte pas les chances de l'autre.

Trois événements indépendants. Les trois événements A, B, C sont dit *mutuellement* indépendants si, et seulement si, ils satisfont les *deux* conditions suivantes :

$$\Pr\{AB\} = \Pr\{A\}\Pr\{B\},\ \Pr\{AC\} = \Pr\{A\}\Pr\{C\},\ \Pr\{BC\} = \Pr\{B\}\Pr\{C\}, \tag{1.15a}$$

$$\Pr\{ABC\} = \Pr\{A\}\Pr\{B\}\Pr\{C\}\,. \tag{1.15b}$$

La première condition définit seulement l'indépendance *deux à deux* des événements, la seconde condition ajoutée à l'indépendance deux à deux permet d'écrire :

$$\Pr\{AB\}\Pr\{C\} = \Pr\{AC\}\Pr\{B\} = \Pr\{BC\}\Pr\{A\} = \Pr\{A\}\Pr\{B\}\Pr\{C\}\,, \tag{1.16}$$

qui exprime qu'un événement quelconque est aussi indépendant de la réalisation simultanée des deux autres. Il est facile de montrer qu'il est aussi indépendant de la réalisation de l'un ou de l'autre des événements restants. Montrons, par exemple, que $\Pr\{A(B \cup C)\} = \Pr\{A\}\Pr\{B \cup C\}$.

On a $\Pr\{A(B \cup C)\} = \Pr\{AB\} + \Pr\{AC\} - \Pr\{ABC\} = \Pr\{A\}(\Pr\{B\} + \Pr\{C\} - \Pr\{BC\}) = \Pr\{A\}\Pr\{B \cup C\}$.

Il n'y a aucune raison pour que des événements deux à deux indépendants soient mutuellement indépendants. Un exemple où trois événements sont deux à deux indépendants mais non mutuellement indépendants est donné sur la figure 1.1.

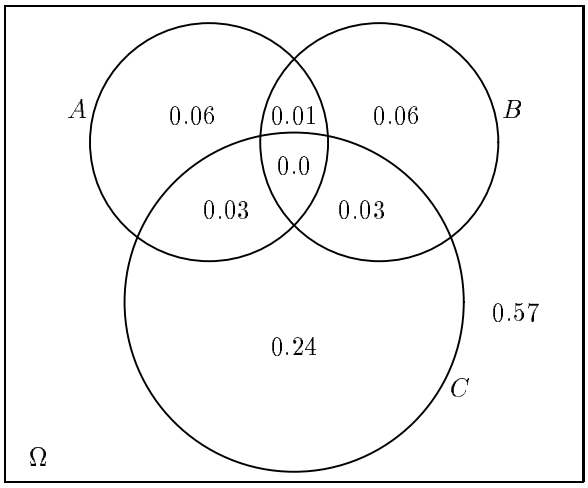


FIG. 1.1: *Les trois événements A, B, C sont* deux à deux *indépendants, mais ils ne sont pas* mutuellement *indépendants. On a* $\Pr\{A\} = 0.1$, $\Pr\{B\} = 0.1$ *et* $\Pr\{C\} = 0.3$, *les probabilités des huits parties définies par les événements A, B et C sont indiquées sur la figure. On vérifie que* $\Pr\{AB\} = 0.01 = \Pr\{A\}\Pr\{B\}$, $\Pr\{AC\} = 0.03 = \Pr\{A\}\Pr\{C\}$ *et* $\Pr\{BC\} = 0.03 = \Pr\{B\}\Pr\{C\}$ *mais on a pas* $\Pr\{ABC\} = \Pr\{A\}\Pr\{B\}\Pr\{C\}$.

Finalement on définit l'indépendance mutuelle d'un nombre quelconque d'événements.

Définition 1.1. Les n événements $A_1, \ldots, A_n$ seront dit *mutuellement* indépendants, si, et seulement si, pour toute combinaison $(i_1, i_2, \ldots, i_k)$ de k indices $(1 < k \leq n)$ extraite de la suite $(1, 2, \ldots, n)$, on a :

$$\Pr\{A_{i_1} A_{i_2} \cdots A_{i_k}\} = \Pr\{A_{i_1}\}\Pr\{A_{i_2}\}\cdots\Pr\{A_{i_k}\}\,. \tag{1.17}$$

Ceci exprime que les événements sont k à k indépendants $(1 < k \leq n)$, c'est-à-dire :

$$\begin{aligned}
\Pr\{A_{i_1} A_{i_2}\} &= \Pr\{A_{i_1}\}\Pr\{A_{i_2}\}\,,\\
\Pr\{A_{i_1} A_{i_2} A_{i_3}\} &= \Pr\{A_{i_1}\}\Pr\{A_{i_2}\}\Pr\{A_{i_3}\}\,,\\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\\
\Pr\{A_1 A_2 \ldots A_n\} &= \Pr\{A_1\}\Pr\{A_2\}\ldots\Pr\{A_n\}\,.
\end{aligned}$$

L'exemple ci-dessous montre que la dernière condition n'implique pas nécessairement les précédentes.

▶**Exemple 1.7.** Dans le jeu de pile ou face $\Omega = [P, F]^3$ où l'on jette 3 pièces de

monnaies, on considère les événements :

$$A = \{PPP, PPF, PFP, PFF\}\,,$$
$$B = \{PPP, PPF, PFP, FPP\}\,,$$
$$C = \{PPP, FPF, FFP, FFF\}\,.$$

Si le jeu est muni de la probabilité uniforme (jeu équitable), on a bien $\Pr\{A,B,C\} = \Pr\{A\}\Pr\{B\}\Pr\{C\} = \frac{1}{8}$, mais $\Pr\{A,B\} = \frac{3}{8}$ n'est pas égal à $\Pr\{A\}\Pr\{B\} = \frac{1}{4}$ et les événements ne sont pas indépendants.

1.4 Exercices

Exercice 1.1. *Règle de de Morgan.* Démontrer la règle de de Morgan (1.4) en notant bien les hypothèses auxquelles il est nécessaire de faire appel. On considérera pour cela l'événement $C = A^c \cap B^c$. Démontrer également que les opérations $\cup$ et $\cap$ sont idempotentes, c'est-à-dire que $A \cup A = A \cap A = A$.

Exercice 1.2. Soit Ω un ensemble des épreuves formé de n événements élémentaires : $\{\omega_1, \ldots, \omega_n\}$. On désigne par T_n le nombre de systèmes complets différents qu'il est possible de construire sur Ω. On ne considérera pas comme étant différents deux événements qui ne diffèrent que par l'ordre de leurs éléments.

Que valent T_1, T_2, T_3, T_4, T_5 et T_6 ?

Montrer que $T_{n+1} = 1 + \sum_{k=1}^{n} C_n^k T_k$. Vérifier que $T_{10} = 115\,975$.

Finalement montrer que :

$$1 + \sum_{k=1}^{\infty} \frac{T_k}{n!} x^k = e^{e^x - 1}\,.$$

Exercice 1.3. Si les événements A_k, $(k = 1, \ldots, n)$ appartiennent à une algèbre $\mathcal{A}$, montrer que l'événement $\cup_{k=1}^{n} A_k$ appartient aussi à $\mathcal{A}$.

Démontrer l'équation (1.8), c'est-à-dire : si les événements A_k, $(k = 1, \ldots, \infty)$ appartiennent à une tribu $\mathfrak{B}$, alors l'événement $\cap_{k=1}^{\infty} A_k$ appartient aussi à la tribu.

Exercice 1.4. Soit un ensemble des épreuves Ω contenant un nombre fini n d'événements élémentaires. Montrer que le nombre d'événements de la plus petite algèbre sur Ω est égal à 2^n. On note 2^Ω cette algèbre.

Exercice 1.5. Par définition une suite d'événements $\{D_k\}$ est dite *décroissante* si : $\forall n;\ D_{n+1} \subset D_n$. On appelle limite de cette suite l'ensemble : $\lim_{k\to\infty} D_k \stackrel{\text{def}}{=} \bigcap_{k=1}^{\infty} D_k$. Montrer que, si $\{A_k\}$ est une suite d'événements *disjoints*, on peut alors remplacer l'axiome **P2** par les deux axiomes équivalents suivants :

P2a : $\Pr\Big\{\bigcup_{k=1}^{n} A_k\Big\} = \sum_{k=1}^{n} \Pr\{A_k\}\,,$ (*additivité finie*),

P2b : $\lim_{n\to\infty} D_n = \varnothing \Rightarrow \lim_{n\to\infty} \Pr\{D_n\} = 0\,,$ (*continuité au vide.*)

Exercice 1.6. *Convexité de* Pr. Montrer que si les événements A_j sont quelconques (c'est-à-dire non nécessairement disjoints), on a :

$$\Pr\Big\{\bigcup_j A_j\Big\} = \Pr\{A_1\} + \Pr\{A_1^c A_2\} + \Pr\{A_1^c A_2^c A_3\} + \cdots \le \sum_j \Pr\{A_j\}\,.$$

Exercice 1.7. Soient des événements $A_1, A_2, \ldots, A_n$ et une mesure de probabilité Pr. Montrer que l'on peut écrire :

$$\Pr\{A_1 A_2 \cdots A_n\} = \Pr\{A_1\}\Pr\{A_2|A_1\}\Pr\{A_3|A_1A_2\}\cdots\Pr\{A_n|A_1A_2\cdots A_{n-1}\}\,.$$

Exercice 1.8. Deux événements incompatibles peuvent-ils être indépendants ?

Exercice 1.9. Montrer que, si A et B sont deux événements indépendants, alors A et B^c sont indépendants ainsi que A^c et B et finalement A^c et B^c.

Exercice 1.10. On définit l'indépendance de deux événements A et B de la façon suivante : A et B seront dit indépendants si $\Pr\{A|B\} = \Pr\{A|B^c\}$. A quelles conditions cette définition est-elle équivalente à celle donnée par la formule (1.14) ?

Chapitre 2

Variables aléatoires.

La notion de variable aléatoire permet de créer une relation entre un espace probabilisé et un espace mesurable. La mesure de probabilité qui fait défaut à l'espace mesurable est définie grâce au lien établi entre les deux espaces. Par ce biais, l'espace qui n'était que mesurable devient lui aussi probabilisé.

L'espace mesurable est presque toujours l'espace arithmétique $\mathbb{R}$ ou $\mathbb{R}^n$. Dans ce cas la variable aléatoire présente le grand intérêt pratique d'associer des nombres à une expérience soumise au hasard dont les issues sont abstraites. Pour un expérimentateur l'espace probabilisé est le phénomène étudié et l'espace mesurable l'ensemble de tous les résultats chiffrés des expériences portant sur le phénomène en question.

2.1 Une variable aléatoire.

Considérons un espace probabilisé, c'est-à-dire un triplet formé d'un ensemble des épreuves Ω, d'une tribu d'événements $\mathfrak{B}$ et d'une mesure de probabilité Pr. Soit ω un événement élémentaire de Ω, associons-lui un nombre réel X à l'aide d'une application ξ.

$$\xi : \Omega \mapsto \mathbb{R} \,. \tag{2.1}$$

Afin d'alléger l'écriture, on notera de façon identique l'application et le réel qui en est le résultat, soit : $X = X(\omega)$.

L'espace $\mathbb{R}$ est muni d'une tribu qui contient les intervalles du type : $]-\infty, x]$. Pour que l'application X soit une variable aléatoire il faut que l'image inverse des intervalles $]-\infty, x]$ soit un événement. Soit A_x cette image inverse, par définition A_x est l'ensemble des ω dont l'image par X appartient à l'intervalle en question :

$$A_x = X^{-1}(]-\infty, x]) \stackrel{\text{def}}{=} \{\omega \mid X(\omega) \in]-\infty, x]\} \,. \tag{2.2}$$

Ainsi X est une *variable aléatoire* si, et seulement si, A_x est un événement, c'est-à-dire : $A_x \in \mathfrak{B}$.

▶**Exemple 2.1.** *Indicatrice d'un événement.* L'indicatrice $\mathbf{1}_A$ d'un événement A

est une variable aléatoire qui vaut 1 si A est réalisé et 0 dans le cas contraire :

$$\mathbf{1}_A(\omega) = \begin{cases} 1 & \text{si } \omega \in A\,; \\ 0 & \text{si } \omega \notin A\,. \end{cases} \tag{2.3}$$

L'indicatrice de A est clairement une variable aléatoire car $\varnothing, A^c$ et Ω sont des événements.

▶**Exemple 2.2.** *Nombres pseudo-aléatoires.* Dans une certaine mesure on peut considérer qu'un programme censé fournir des nombres au hasard se comporte comme une variable aléatoire X. Les nombres qu'il délivre sont les valeurs x prises par cette variable.

L'association, par la pensée, d'une variable aléatoire avec un programme permet parfois de clarifier certaines notions du calcul des probabilités. Par exemple, nous verrons plus bas l'équation : $\Pr\{X \leq x\}$, qui pourra s'interpréter comme la probabilité pour que le programme X fournisse des nombres ne dépassant pas la valeur x. Une image semblable se révélera également utile lorsque l'on étudiera au chapitre 7 la convergence d'une variable aléatoire vers une autre.

Loi d'une variable aléatoire.

Si X est une variable aléatoire il est alors possible d'associer à l'intervalle $]-\infty, x]$ une probabilité qui, par définition, sera égale à celle associée à A_x. Plus précisément :

Définition 2.1. On appelle *loi* d'une variable aléatoire X, la probabilité image de Pr par X, on note $\Pr^X$ cette probabilité. La loi d'une variable aléatoire X n'est autre que la mesure de probabilité induite par X sur $\mathbb{R}$ (ou $\mathbb{R}^n$). On a donc :

$$\Pr^X\{X(\omega) \in]-\infty, x]\} \stackrel{\text{def}}{=} \Pr\{A_x\}\,. \tag{2.4}$$

▶**Exemple 2.3.** *Loi de Bernoulli.* La loi suivie par l'indicatrice de A est appelée « loi de Bernoulli ». On note souvent $\mathcal{B}(1,p)$ l'ensemble des variables aléatoires qui suivent la loi de Bernoulli, p désigne la probabilité de A.

2.1.1 Fonction de répartition.

Soit X une variable aléatoire à valeurs réelles $X \in \mathbb{R}$. La fonction de répartition de X est égale à la probabilité de l'intervalle $]-\infty, x]$ envisagée comme fonction de x, soit :

$$\begin{aligned} F_X(x) &= \Pr^X\{X \in]-\infty, x]\}\,, \\ &= \Pr\{X^{-1}(]-\infty, x])\}\,. \end{aligned}$$

Ainsi l'information sur la loi $\Pr^X$ est entièrement contenue dans sa fonction de répartition. On dira que l'on connaît la loi suivie par la variable aléatoire X si l'on connaît sa fonction de répartition.

Le plus souvent on supprimera l'indice ^{X} et on notera simplement Pr la mesure de probabilité induite par X sur $\mathbb{R}$. On dira aussi que la fonction de répartition est « la probabilité pour que la variable aléatoire X soit inférieure ou égale à x », au lieu de « la probabilité pour que le résultat de l'application X soit inférieure ou égale à x. » Cet abus de langage ne prête généralement pas à

confusion, on notera d'ailleurs : $F_X(x) = \Pr\{X \leq x\}$. Enfin, lorsqu'il sera clair que la fonction de répartition F_X fera référence à la variable aléatoire X, on la notera simplement F.

Avec toutes ces conventions, la définition de la fonction de répartition peut s'écrire de façon impropre, mais bien pratique :

$$F(x) = \Pr\{X \leq x\}\,. \tag{2.5}$$

La fonction de répartition d'une variable aléatoire X s'interpète alors comme la probabilité pour que X ne dépasse pas un certain seuil x. L'ensemble $\mathcal{X}$ de tous les résultats possibles de X est appellé *le domaine de définition* de X.

Remarque 2.1. Cette définition de la fonction de répartition correspond à la convention anglo-saxonne, la convention française serait plutôt $F_X(x) = \Pr\{X < x\}$.

▶**Exemple 2.4.** *Fonction de répartition de la Loi de Laplace-Gauss.* La fonction de répartition, que l'on note Φ, d'une variable aléatoire suivant la loi de Laplace-Gauss, est donnée par l'expression :

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}t^2}\, dt. \tag{2.6}$$

La loi de Laplace-Gauss est également appelée *Loi normale*. Le graphe de la fonction de répartition de cette loi est donnée par la figure 2.1

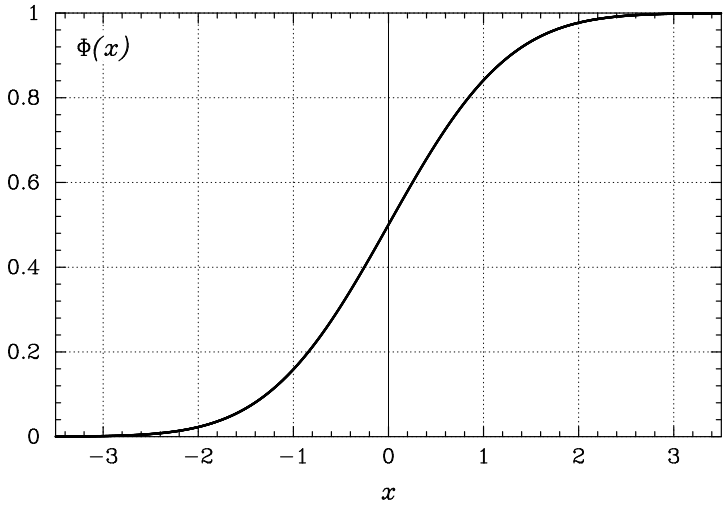


FIG. 2.1: *Fonction de répartition de la Loi de Laplace-Gauss (ou loi normale). Il s'agit ici du graphe de la loi normale réduite.*

2.1.2 Probabilité attachée à un intervalle.

Rapellons qu'un *intervalle* de $\mathbb{R}$ est tout ensemble de nombres réels de la forme :

$$\{x \mid a < x < b\},\ \{x \mid a \leq x < b\},\ \{x \mid a < x \leq b\} \text{ ou } \{x \mid a \leq x \leq b\}. \tag{2.7}$$

Nous les notons respectivement : $]a,b[$, $[a,b[$, $]a,b]$ et $[a,b]$. Les anglo-saxons les notent souvent : (a,b), $[a,b)$, $(a,b]$ et $[a,b]$. Dans ces expressions les symboles a et b représentent aussi bien des nombres réels finis ($a,b \in \mathbb{R}$) que des nombres infinis ($a,b \in \overline{\mathbb{R}}$). A ces intervalles sont attachées les probabilités suivantes :

$$\Pr\{a < X \leq b\} = F(b) - F(a)\,; \tag{2.8a}$$
$$\Pr\{a \leq X \leq b\} = F(b) - F(a) + \Pr\{X = a\}\,; \tag{2.8b}$$
$$\Pr\{a \leq X < b\} = F(b) - F(a) + \Pr\{X = a\} - \Pr\{X = b\}\,; \tag{2.8c}$$
$$\Pr\{a < X < b\} = F(b) - F(a) - \Pr\{X = b\}\,. \tag{2.8d}$$

Démonstration. Définissons les événements : $A = \{X \leq a\}$ et $B = \{X \leq b\}$. Notons que si $a < b$ (comme nous le supposons), on a $A \subset B$. Les événements B^c, BA^c et A forment alors une partition de Ω il vient : $1 = \Pr\{B^c\} + \Pr\{BA^c\} + \Pr\{A\}$, d'où on tire $\Pr\{BA^c\} = \Pr\{B\} - \Pr\{A\}$. Il s'ensuit :

$$\Pr\{a < X \leq b\} = \Pr\{BA^c\} = \Pr\{B\} - \Pr\{A\} = F(b) - F(a)\,. \tag{2.9a}$$

Notons que $\Pr\{X \leq b\} = \Pr\{X < b\} + \Pr\{X = b\}$ et que $\Pr\{a \leq X\} = \Pr\{a < X\} + \Pr\{X = a\}$ d'où on tire immédiatement les autres résultats. □

▶**Exemple 2.5.** *Intervalles de la Loi de Laplace-Gauss.* Il n'y a pas lieu de distinguer entre les intervalles définis en (2.7) car $\Pr\{X = x\} = 0$. La probabilité pour qu'une variable aléatoire X suivant la loi de Laplace-Gauss, soit comprise entre -1 et 1 est égale à $\Phi(1) - \Phi(-1)$ et vaut $\approx 0.8413 - 0.1587 = 0.6826$.

2.1.3 Propriétés de la fonction de répartition.

Une fonction quelconque n'est en général pas la fonction de répartition d'une variable aléatoire. On montre que F ne peut être une fonction de répartition que si, et seulement si, elle possède les propriétés suivantes :

1. ses valeurs $F(x)$ sont toujours comprise entre 0 et 1 ;
2. elle est croissante : $x_2 \geq x_1 \iff F(x_2) \geq F(x_1)$;
3. elle est continue à droite en tout point de son domaine de définition, c'est-à-dire : $\forall x \in \mathcal{X}, \forall \epsilon > 0$; $\lim_{\epsilon \to 0} F(x+\epsilon) = F(x)$, ce que l'on note $F(x^+) = F(x)$;
4. la propriété 1 nous dit qu'elle est bornée sur $\mathcal{X} \subseteq \mathbb{R}$ et on adopte, s'il y a lieu, la convention : $F(-\infty) = 0$ et $F(+\infty) = 1$.

La première propriété résulte directement du fait que $F(x)$ est une probabilité. La deuxième et la troisième également, en effet calculons la probabilité pour que la variable aléatoire X soit comprise dans l'intervalle $]x_1, x_2]$:

$$\Pr\{x_1 < X \leq x_2\} = F(x_2) - F(x_1). \tag{2.10}$$

Une probabilité étant, par définition, un nombre non-négatif, on a nécessairement $F(x_1) \leq F(x_2)$, ce qui montre que F est croissante. Elle n'est cependant

pas strictement croissante, c'est-à-dire qu'elle peut présenter des plateaux. La troisième propriété s'obtient en posant $x_2 = x_1 + \epsilon$ et par passage à la limite :

$$\begin{aligned} \lim_{\epsilon \to 0} \Pr\{x < X \leq x + \epsilon\} &= F(x + \epsilon) - F(x)\,, \\ \Pr\{\varnothing\} &= F(x^+) - F(x)\,, \\ &= 0\,. \end{aligned}$$

En revanche F n'est pas nécessairement continue à gauche, on a :

$$\begin{aligned} \lim_{\epsilon \to 0} \Pr\{x - \epsilon < X \leq x\} &= F(x) - F(x - \epsilon)\,, \\ \Pr\{X = x\} &= F(x) - F(x^-)\,. \end{aligned}$$

Ce qui peut s'écrire sous la forme :

$$\Pr\{X = x\} = F(x^+) - F(x^-)\,, \tag{2.11}$$

qui présente l'avantage d'être valable quelle que soit la convention (anglo-saxonne ou française) choisie pour définir la fonction de répartition. Cette formule implique que F est continue en x si, et seulement si, $\Pr\{X = x\} = 0$. La quantité $F(x_i^+) - F(x_i^-)$ représente le saut de la fonction en x_i. Dans le cas où $F(x^+) \neq F(x^-)$, on conviendra de représenter $F(x^+)$ par un point épais sur le graphe de la fonction F (voir figure 2.2).

Les conditions 1–3 sont donc nécessaires pour que F soit un fonction de répartition, nous admettrons le théorème suivant qui dit qu'elles sont aussi suffisantes.

Théorème 2.1. *Pour que la fonction F soit la fonction de répartition d'une variable aléatoire quelconque il faut et il suffit que F soit une fonction monotone croissante, continue à droite et admette les limites 0 en $-\infty$ et 1 en $+\infty$.*

Finalement F, comme toute fonction monotone, n'admet qu'un ensemble dénombrable de points de discontinuité (ensemble qui peut d'ailleurs être dense).

2.1.4 Différents types de fonctions de répartition.

Il n'y a que trois types de fonctions qui peuvent être des fonctions de répartition : les fonctions discontinues dites *en escaliers*; et les fonctions continues qui se scindent en fonctions *absolument continues* et fonctions *singulièrement continues*. A ces trois types de fonctions sont attachés trois types de variables aléatoires : les variables *discrètes*; les variables *absolument continues* et les variables *continues singulières*. Seuls les deux premiers types de variables aléatoires présentent un intérêt pratique.

On montre qu'une fonction de répartition quelconque est la somme d'une fonction F^a, F^{ac}, F^{sc} de chacun des types, on a :

$$F(x) = a_1 F^a(x) + a_2 F^{ac}(x) + a_3 F^{sc}(x)\,. \tag{2.12}$$

Une variable aléatoire est donc discrète si $a_1 \neq 0\,, a_2 = a_3 = 0$, absolument continue si $a_2 \neq 0\,, a_1 = a_3 = 0$ et elle sera dite de type *mixte* si a_1 et a_2 ne sont pas nuls alors que a_3 est nul.

Les variables aléatoires discrètes.

Ce sont les variables dont la fonction de répartition F est en escaliers (on dit encore *réglée*), c'est-à-dire : discontinue et constante entre les points de discontinuité. La probabilité attachée à un point est presque partout nulle sauf aux points de discontinuité de F. Les points de discontinuité étant dénombrables les valeurs prise par cette variable aléatoire sont eux aussi dénombrables.

▶**Exemple 2.6.** *Loi de Poisson* Une variable aléatoire X suit la loi de Poisson si elle possède une probabilité non nulle pour tout x entier positif ou nul et si elle possède la fonction de répartition suivante (valable pour $\mu > 0$) :

$$F(x) = \sum_{k=0}^{\lfloor x \rfloor} \frac{\mu^k}{k!} e^{-\mu}, \quad x \geq 0 . \tag{2.13}$$

L'expression $\lfloor x \rfloor$ désigne le plus grand entier $\leq$ à x. Le graphe de cette fonction est représenté sur la figure 2.2.

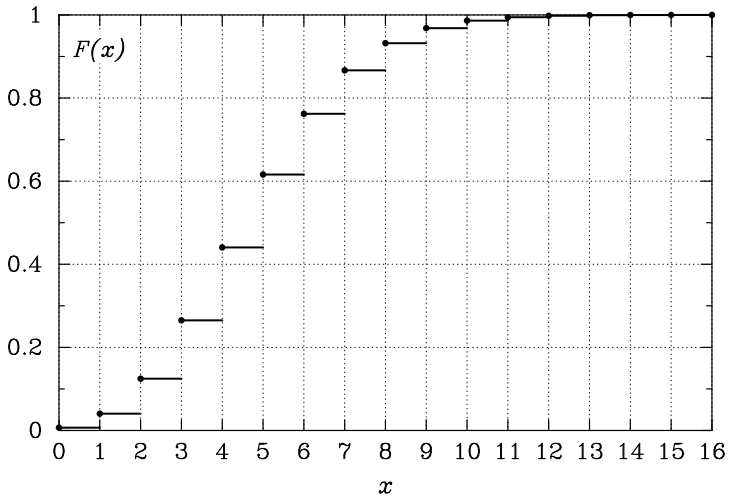


FIG. 2.2: *Exemple de fonction de répartition d'une variable aléatoire discrète. Il s'agit ici de la fonction de répartition d'une variable aléatoire X suivant la loi de Poisson de paramètre $\mu = 5$. Avec la définition $F(x) = \Pr\{X \leq x\}$, cette fonction est alors continue à droite.*

Les variables absolument continues.

Une fonction F est absolument continue si, quel que soit le nombre $\epsilon > 0$, il existe un nombre η tel que :

$$\forall n, \sum_{k=1}^{n} |b_k - a_k| < \eta \Rightarrow \sum_{k=1}^{n} |F(b_k) - F(a_k)| < \epsilon ,$$

pour tout système d'intervalles disjoints d'extrémités a_k et b_k.

La différence entre la définition de la continuité et de l'absolue continuité porte sur l'introduction de la somme $\sum_{k=1}^{n}$, il est alors évident qu'une fonction absolument continue est continue mais la réciproque n'est pas vraie. Il est plus

intuitif de caractériser les fonctions absolument continues à l'aide de la propriété équivalente suivante : une fonction est absolument continue si, et seulement si, elle est presque-partout dérivable et égale à l'intégrale indéfinie de sa dérivée. Suivant cette définition F est absolument continue si, et seulement si :

$$F(x) = \int_{-\infty}^{x} f(t)\,dt\,, \quad f(t) = F'(t) \textit{ presque-partout}. \tag{2.14}$$

Il est difficile de donner, à l'aide de formules simples, un exemple de fonction singulièrement continue. Ces fonctions ne présentent pas, a l'heure actuelle, d'application pratique c'est pourquoi nous ne les considérerons pas plus avant et quand on parlera, dans ce texte, de fonctions ou de variables aléatoires « continues » il faudra entendre « absolument continues ».

▶**Exemple 2.7.** *Loi exponentielle.* La variable aléatoire X suit la loi exponentielle si, quel que soit $\lambda > 0$, elle possède la fonction de répartition suivante :

$$F(x) = \begin{cases} 1 - e^{-\lambda x} & \text{si } x \geq 0\,; \\ 0 & \text{si } x < 0\,. \end{cases} \tag{2.15}$$

La fonction F est dérivable sauf en $x = 0$ et $F(x) = \int_0^x \lambda e^{-\lambda t}\,dt$ pour $x \geq 0$, c'est donc une fonction de répartition absolument continue. Le graphe de cette fonction est représenté sur la figure 2.3.

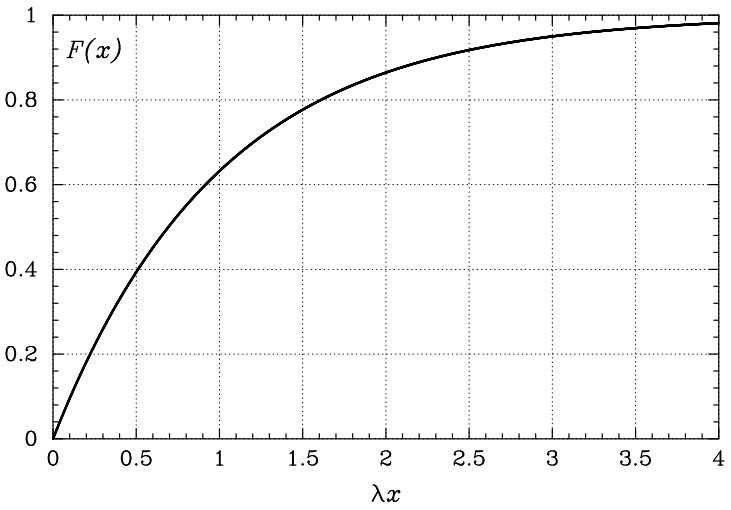


FIG. 2.3: *Exemple de fonction de répartition d'une variable aléatoire absolument continue. Il s'agit ici de la fonction de répartition d'une variable aléatoire X suivant la loi exponentielle de paramètre $\lambda > 0$.*

2.1.5 Densité de probabilité.

Considérons la probabilité pour qu'une variable aléatoire X soit comprise entre x et $x + \Delta x$, d'après les résultats des équations (2.8) on a :

$$\Pr\{x < X \leq x + \Delta x\} = F(x + \Delta x) - F(x)\,.$$

Si F est absolument continue on a : $F(x+\Delta x) - F(x) = \int_x^{x+\Delta x} f(t)\,dt$, où f est la dérivée de F, cette dérivée existe pour presque tous les x du domaine de définition de X. L'expression précédente s'écrit alors :

$$\Pr\{x < X \leq x + \Delta x\} = f(x)\Delta x + o(\Delta x)\,. \tag{2.16}$$

En faisant tendre Δx vers 0 il vient :

$$\lim_{\Delta x \to 0} \frac{\Pr\{x < X \leq x + \Delta x\}}{\Delta x} = f(x)\,. \tag{2.17}$$

Ce qui montre que $f(x)$ peut s'interpréter comme la densité (linéaire) de probabilité au point x.

Définition 2.2. Par définition, la densité de probabilité d'une variable aléatoire X absolument continue est égale à la dérivée de sa fonction de répartition aux points où cette dérivée existe :

$$f(x) = \frac{dF(x)}{dx} \; \textit{presque-partout}\,. \tag{2.18}$$

▶**Exemple 2.8.** *Densité de probabilité de la loi normale.* La fonction de répartition Φ d'une variable aléatoire X suivant la loi normale est donnée par l'expression (2.6). La fonction Φ est absolument continue sur $\mathbb{R}$ et sa dérivée est définie pour tout $x \in \mathbb{R}$. La densité de probabilité de X est alors donnée par l'expression :

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\{-\tfrac{1}{2}x^2\}\,,$$

son graphe est donnée sur la figure 2.4

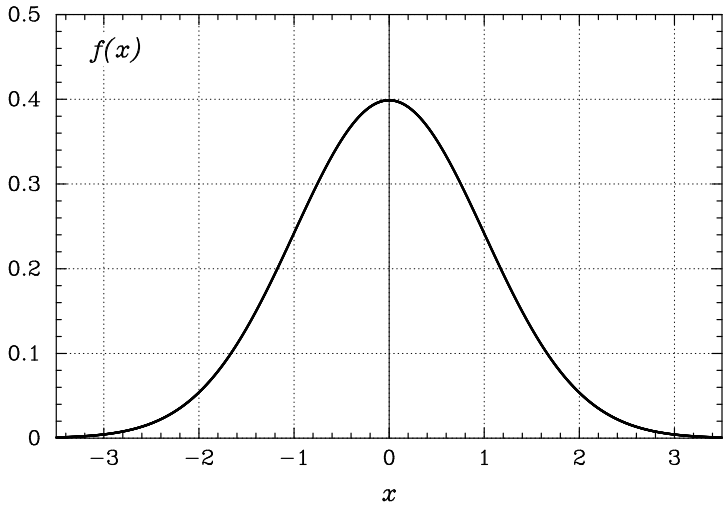


FIG. 2.4: *Densité de probabilité de la Loi de Laplace-Gauss (ou loi normale).*

Remarque 2.2. En toute rigueur la densité de probabilité n'est définie que pour les variables aléatoires absolument continues. Cependant on trouve dans certains ouvrages (surtout orientés vers les applications pratiques), une extension de la notion de densité de probabilité appliquée aux variables discrètes et mixtes.

Dans cette acceptation, la dérivée (2.18) est à prendre au sens des distributions. Les fonctions de répartition ne comportant qu'un nombre au plus dénombrable de discontinuités, on a alors la formule :

$$F' = \{F\}' + \sum_{i \in I} p_i \delta_{x_i} \,, \tag{2.19}$$

où $\{F\}'$ est la dérivée de F au sens des fonctions, $p_i = F(x_i^+) - F(x_i^-)$ est la valeur du saut de la fonction F en ses discontinuités situées en x_i et $\delta_{x_i}(x) = \delta(x - x_i)$ est une translatée de la distribution de Dirac. Sur la figure 2.5 la « densité » de la loi de Poisson de paramètre $\mu = 5$ est tracée sous la forme d'un « diagramme en bâtons. » Dans ce diagramme les sauts de F sont représentés par des traits verticaux, ils corespondent aux distributions δ de l'équation (2.19) ci-dessus.

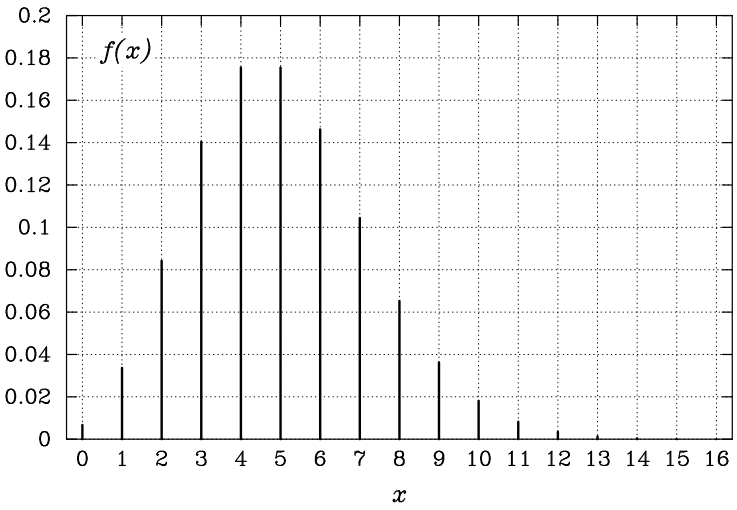


FIG. 2.5: « *Densité* » *de probabilité de la Loi de Poisson pour* $\mu = 5$.

Il faut néanmoins noter que l'introduction des distributions n'est pas nécessaire en théorie des probabilités parce que l'outil fondamental est la fonction de répartition pas la densité de probabilité. L'usage de cette densité impropre constitue cependant un moyen commode d'acquérir une vision intuitive du lieu où se trouvent les valeurs « les plus probables » de la variable aléatoire.

2.1.6 Propriétés de la densité de probabilité.

On sait que la fonction de répartition s'exprime ainsi à l'aide de la densité de probabilité :

$$F(x) = \int_{-\infty}^{x} f(t)\, dt \,, \tag{2.20}$$

et que f est presque-partout égale à la dérivée de F. De plus une densité de probabilité est normalisée et positive :

$$\int_{-\infty}^{+\infty} f(t)\, dt = F(+\infty) = 1 \,, \quad f(x) \geq 0 \,. \tag{2.21}$$

Réciproquement, toute fonction *mesurable* (approximativement cela veut dire intégrable), *normalisée* et *positive* est la densité de probabilité d'une certaine variable aléatoire.

2.2 Caractéristiques numériques des lois 1D.

Une loi quelconque est entièrement décrite par sa fonction de répartition ou sa densité de probabilité, mais cette information est souvent trop riche, pour être facilement appréhendée, et l'on souhaite alors caractériser la loi par un ensemble restreint de paramètres. Nous allons maintenant définir certains de ces paramètres.

2.2.1 Le mode.

Un mode m est une valeur au voisinage de laquelle la probabilité d'obtenir m à Δx près est maximum. Plus précisément, m est un mode s'il existe un $\Delta x_0 > 0$ tel que pour tout $\Delta x \neq 0$ compris entre 0 et Δx_0 on ait :

$$\forall x \in \mathcal{X},\ \Pr\{X \in]x - \Delta x, x + \Delta x]\} < \Pr\{X \in]m - \Delta x, m + \Delta x]\}\,, \qquad (2.22)$$

pour tout x différent de m. Intuitivement le mode est la valeur que l'on « rencontre » le plus souvent(à Δx près pour les variables continues), c'est d'une certaine façon la valeur « à la mode ».

►**Exemple 2.9.** La loi de Poisson de paramètre $\mu = 5$ possède un mode en $X = 4$ et un autre en $X = 5$ (voir figure 2.5).

S'il n'y a qu'un seul mode, nous dirons que c'est *le mode* de la loi. En revanche une loi quelconque peut ne pas posséder de mode.

►**Exemple 2.10.** Suivant la définition (2.22) la loi exponentielle ne possède pas de mode. Par extension, elle possède un mode en 0^+, on peut alors dire qu'elle possède un mode en zéro. La loi uniforme ne possède aucun mode.

Une valeur m est un *mode local* si l'équation (2.22) est satisfaite localement, c'est-à-dire non pas pour tout $x \in \mathcal{X}$ mais pour les x d'un intervalle $I \subseteq \mathcal{X}$ contenant m. Un mode est évidemment un mode local et on dira qu'une loi est *unimodale* si elle ne possède qu'un mode local. On dira qu'elle est *multimodale* si elle possède au moins deux modes locaux.

►**Exemple 2.11.** La loi normale est unimodale et son mode vaut zéro (voir figure 2.4, page 22). Si la loi est discrète, il existe un mode local en x_i si, et seulement si : $\Pr\{X = x_i\} > \Pr\{X = x_{i-1}\}$, et $\Pr\{X = x_i\} > \Pr\{X = x_{i+1}\}$.

Lorsque la loi dont on cherche le mode est continue et possède une densité f alors le mode de la loi est le maximum de f. Si f' et f'' existent, alors une condition suffisante pour que m soit un mode local est que :

$$m \in \mathcal{X}, \quad f'(m) = 0, \quad f''(m) < 0\,, \qquad (2.23)$$

$$\text{ou bien :} \quad m \in \mathcal{X}, \quad F''(m) = 0, \quad F'''(m) < 0\,. \qquad (2.24)$$

Ce mode local correspond à un point d'inflexion de la fonction de répartition.

Le mode d'une loi inconnue mais dont on a observé quelques réalisations, n'est pas une valeur facile à estimer, c'est pourquoi on lui préfère d'autres paramètres pour caractériser la loi.

2.2.2 Les moments.

Les moments non-centrés. Les moments non-centrés μ'_k d'ordre k d'une variable aléatoire X suivant la loi F sont définis par les intégrales suivantes :

$$\mu'_k = \int_{\mathcal{X}} x^k \, dF \,, \; k \in \mathbb{N} \,. \tag{2.25}$$

Dans cette expression l'intégrale est l'intégrale de Stieltjes, elle porte sur $\mathcal{X}$ le domaine de définition de X. L'intégrale de Stieltjes tient compte des discontinuités éventuelles de F. On la calcule en écrivant que la fonction de répartition d'une variable mixte est la somme d'une fonction en escalier et d'une fonction absolument continue [voir équation (2.12)]. On envisage les deux cas séparément ci-dessous :

- Si X est une variable aléatoire discrète prenant les valeurs $(x_1, x_2, \ldots)$, sa fonction de répartition F est « en escaliers » et l'intégrale de Stieltjes s'écrit :

$$\begin{aligned} \int_{\mathcal{X}} x^k \, dF &= \sum_i x_i^k (F^+(x_i) - F^-(x_i)) \,, \\ &= \sum_i x_i^k \Pr\{X = x_i\} \,. \end{aligned} \tag{2.26}$$

  L'intégrale est définie si la série du membre de droite est absolument convergente c'est-à-dire, en posant $p_i = \Pr\{X = x_i\}$, si : $\sum_i |x^k| p_i < \infty$.

  ▶**Exemple 2.12.** *Loi de Bernoulli.* Les moments non-centrés d'une variable aléatoire de Bernoulli : $X \in \mathcal{B}(1, p)$ sont tous égaux à p, en effet :

$$\int x^k \, dF = 0^k (1 - p) + 1^k p = p \,.$$

- Si X est une variable aléatoire absolument continue, elle possède alors une densité f et l'intégrale de Stieltjes est identique à l'intégrale de Riemann :

$$\int_{\mathcal{X}} x^k \, dF = \int_{\mathcal{X}} x^k f(x) \, dx \,. \tag{2.27}$$

  Elle est définie si $\int_{\mathcal{X}} |x^k| f(x) \, dx < \infty$.

Le moment non-centré d'ordre 0 existe et est, d'après (2.21), toujours égal à 1. En revanche les moments non-centrés d'ordres supérieurs à 0 peuvent ne pas exister. Si un moment n'existe pas à l'ordre k, il n'existe alors aucun moments à l'ordre $r > k$.

▶**Exemple 2.13.** *La loi de Cauchy.* Une variable aléatoire suit la loi de Cauchy si elle possède la densité :

$$f(x) = \frac{1}{\pi} \frac{1}{1 + x^2} \,.$$

Cette loi n'a aucun moment car $\forall k > 0$ l'intégrale (2.27) diverge.

▶**Exemple 2.14.** *Le jeu de S^t Petersbourg.* On considère un jeu de « pile » ou « face » infini : $\Omega = \{P, F\}^\infty$, p est la probabilité d'obtenir « face ». On distribue le gain $X = p^{-n}$ au joueur qui a obtenu n « face » consécutifs. Si $p = \frac{1}{2}$, par exemple, on aura : $X(\{F\}) = 2$, $X(\{FF\}) = 4$, $X(\{FFF\}) = 8$, etc... Le gain X est nul dans tous les autres cas. Il est facile de voir que cette variable X ne possède pas de moments. Par exemple pour le moment d'ordre $k = 1$ on a : $\sum_{i=1}^{\infty} p^{-n} p^n = 1 + 1 + 1 + \ldots$, la série diverge.

Les moments centrés. On pose en général $\mu = \mu'_1$ et les moments centrés, s'ils existent, sont définis par :

$$\mu_k = \int_{\mathcal{X}} (x-\mu)^k \, dF \,, \; k \in \mathbb{N} \,. \tag{2.28}$$

Un moment centré n'existe que si, et seulement si, le moment non-centré correspondant existe lui-aussi.

En développant $(x-\mu)^k$ dans l'équation (2.28), on trouve les relations entre les moments centrés et non-centrés. On obtient pour les quatre premiers moments :

$$\mu_0 = 1 \qquad \mu'_0 = 1$$
$$\mu_1 = 0 \qquad \mu'_1 = \mu$$
$$\mu_2 = \mu'_2 - \mu^2 \qquad \mu'_2 = \mu_2 + \mu^2 \tag{2.29a}$$
$$\mu_3 = \mu'_3 - 3\mu'_2\mu + 2\mu^3 \qquad \mu'_3 = \mu_3 + 3\mu_2\mu + \mu^3 \tag{2.29b}$$
$$\mu_4 = \mu'_4 - 4\mu'_3\mu + 6\mu'_2\mu^2 - 3\mu^4 \qquad \mu'_4 = \mu_4 + 4\mu_3\mu + 6\mu_2\mu^2 + \mu^4 \tag{2.29c}$$

Parmi tous les moments, on distingue (sous réserve d'existence) :

La moyenne. Le moment μ'_1 reçoit le nom de « *moyenne* » et est généralement noté μ. Par définition on a :

$$\mu = \int_{\mathcal{X}} x \, dF \quad \text{et si } f \text{ existe} \quad \mu = \int_{\mathcal{X}} x f(x) \, dx \tag{2.30}$$

La moyenne μ peut s'interpréter comme l'abscisse du centre de gravité de l'axe des réels ayant $f(x)$ comme densité linéaire de masse.

La variance et l'écart type. Le moment centré μ_2, reçoit le nom de « *variance* », c'est une quantité positive que l'on note généralement σ^2. L'écart type σ est défini comme la racine carrée de la variance. Ainsi défini σ est toujours positif. On a :

$$\sigma^2 = \int_{\mathcal{X}} (x-\mu)^2 \, dF \quad \text{et si } f \text{ existe} \quad \sigma^2 = \int_{\mathcal{X}} (x-\mu)^2 f(x) \, dx \,. \tag{2.31}$$

La variance correspond au moment d'inertie de l'axe réel de densité $f(x)$ calculé autour de la moyenne μ. D'après (2.29a) la variance d'une variable aléatoire X s'exprime en fonction des deux premiers moments non-centrés :

$$\sigma^2 = \mu'_2 - \mu^2 \,. \tag{2.32}$$

Asymétrie et aplatissement. On utilise également les moments centrés d'ordres 3 et 4, μ_3 et μ_4 pour définir le coefficient d'asymétrie γ_1 et le coefficient d'aplatissement γ_2 de la façon suivante :

$$\gamma_1 = \frac{\mu_3}{\mu_2^{3/2}} \,, \quad \gamma_2 = \frac{\mu_4}{\mu_2^2} - 3 \,. \tag{2.33}$$

Le coefficient γ_2 a été défini ainsi, de façon à être nul pour la loi normale. Une variable aléatoire dont le coefficient d'aplatissement γ_2 est positif est dite

hypernormale et son graphe sera plus pointu que celui d'une loi normale de même variance ; par contre si γ_2 est négatif, la loi sera dite *hyponormale* et son graphe sera plus plat que celui d'une loi normale de même variance.

Remarque 2.3. On trouve parfois dans la littérature une autre définition des coefficients d'asymétrie et d'aplatissement, que l'on note β_1 et β_2 et pour lesquels on a les relations :

$$\beta_1 = \gamma_1^2 \,; \quad \beta_2 = \gamma_2 + 3 \,. \tag{2.34}$$

2.2.3 Variable aléatoire centrée et réduite.

Une variable aléatoire de moyenne nulle est dite *centrée*. Une variable aléatoire de moyenne nulle et d'écart type unité est dite *réduite*. La variable aléatoire $X - \mu$ est centrée et la variable aléatoire $(X - \mu)/\sigma$ est réduite comme on le calcule facilement à l'aide des définitions (2.25) et (2.28).

2.2.4 La médiane et les quantiles.

La médiane $x_{0.5}$ est solution de l'équation :

$$F(x_{0.5}) = \tfrac{1}{2} \,. \tag{2.35}$$

Plus généralement le quantile x_α d'ordre α est solution de l'équation :

$$F(x_\alpha) = 1 - \alpha \,, \quad 0 \leq \alpha \leq 1 \,, \tag{2.36}$$

ce qui s'écrit aussi à l'aide de la densité de probabilité :

$$\int_{x_{0.5}}^{\infty} f(t)dt = 0.5 \,, \quad \int_{x_\alpha}^{\infty} f(t)dt = \alpha \,. \tag{2.37}$$

Si la fonction F présente des plateaux, cette dernière équation peut ne pas avoir de solution unique. Par convention on choisira un point particulier du plateau, en général le point milieu ou une des extrémités. Ce cas se présente toujours lorsqu'on a affaire à des variables aléatoires discrètes.

On publie les quantiles d'une loi donnée, dans une table à deux colonnes dont l'une est la valeur choisie α, et l'autre la valeur du quantile correspondant x_α. La table 2.1 est extraite d'une table de quantiles de la loi normale réduite de fonction de répartition donnée par l'équation (2.6).

α	x_α	α	x_α	α	x_α
0.001	3.0902	0.01	2.3263	0.1	1.2816
0.002	2.8782	0.02	2.0537	0.2	0.8416
0.003	2.7478	0.03	1.8808	0.3	0.5244
0.004	2.6521	0.04	1.7507	0.4	0.2533
0.005	2.5758	0.05	1.6449	0.5	0

TAB. 2.1: *Extrait d'une table de quantiles de la loi normale réduite.*

2.3 Lois conditionnelles.

Si l'on ne s'intéresse à la variable aléatoire X que lorsqu'un certain événement A est réalisé, on obtient grâce à la règle des probabilités composées :

$$\Pr\{X \leq x, A\} = \Pr\{A\} \Pr\{X \leq x | A\} \ . \tag{2.38}$$

L'événement A est appelé « *événement conditionnel* ». Par définition, la « *fonction de répartition conditionnelle* », relative à la condition A, est la probabilité pour que X ne dépasse pas le seuil x, sachant que A est réalisé. On note $F_{X|A}$ cette fonction et l'on a :

$$F_{X|A}(x) \equiv \Pr\{X \leq x | A\} = \frac{\Pr\{X \leq x, A\}}{\Pr\{A\}} \ . \tag{2.39}$$

La densité conditionnelle $f_{X|A}$ s'obtient par dérivation de $F_{X|A}$:

$$f_{X|A}(x) = \frac{d}{dx} \Pr\{X \leq x, A\} \, / \, \Pr\{A\} \ .$$

En général A est indépendant de l'événement $\{X \leq x\}$, il vient alors :

$$\frac{d}{dx} \Pr\{X \leq x, A\} = \frac{d}{dx} \Pr\{X \leq x\} = \frac{dF(x)}{dx} \ ,$$

et il vient :

$$f_{X|A}(x) = \begin{cases} (\Pr\{A\})^{-1} f(x) & \text{si } x \in A \ ; \\ 0 & \text{si } x \notin A \ . \end{cases} \tag{2.40}$$

La densité conditionnelle relativement à l'événement A est donc identique à la densité de X dans le domaine où A est réalisé, et on s'assure que son intégrale sur A est bien égale à un en la renormalisant par la constante $(\Pr\{A\})^{-1}$.

2.3.1 Les lois tronquées.

Intéressons-nous à la loi conditionnelle quand l'événement $A = \{a < X \leq b\}$ est réalisé. La fonction de répartition conditionnelle est alors trouvée à l'aide de (2.39) et vaut :

$$F_{X|A}(x | a < X \leq b) = \begin{cases} 0 & \text{si } x \leq a \ ; \\ \dfrac{F_X(x) - F_X(a)}{F_X(b) - F_X(a)} & \text{si } a < x \leq b \ ; \\ 1 & \text{si } x > b \ . \end{cases} \tag{2.41}$$

où F_X est la fonction de répartition de X. Cette loi possède des moments, par exemple, elle a pour moyenne (conditionnelle) :

$$\mu_A = \frac{1}{F_X(b) - F_X(a)} \int_a^b x \, dF_X \ . \tag{2.42}$$

2.3.2 Lois conditionnelles par rapport à un système d'événements.

Supposons que l'événement conditionnel A soit formé d'un système complet d'événements disjoints, en nombre éventuellement infini mais dénombrable : $A_i; i = 1, \ldots, \infty$. Par définition, les événements constituant ce système sont incompatibles et recouvrent tout A. On a donc :

$$\Pr\{X \leq x, A\} = \sum_i \Pr\{X \leq x, A_i\} = \sum_i \Pr\{X \leq x | A_i\} \Pr\{A_i\} \ . \quad (2.43)$$

Conformément à (2.39), on obtient les fonctions de répartition en divisant (normalisant) par $\Pr\{A\}$. Il vient :

$$F_{X|A}(x) = \sum_i F_{X|A_i}(x) \frac{\Pr\{A_i\}}{\Pr\{A\}} = \sum_i F_{X|A_i}(x) \Pr\{A_i|A\} \ . \quad (2.44)$$

Pour x fixé, $F_{X|A_i}(x)$ est une variable aléatoire discrète et l'équation précédente exprime le fait que la fonction de répartition conditionnelle $F_{X|A}$ est la moyenne de cette variable aléatoire, moyenne calculée sur les A_i. Naturellement, cette formule est encore valable si A représente tout l'ensemble Ω. La formule précédente devient alors :

$$F_X(x) = \sum_i F_{X|A_i}(x) \Pr\{A_i\} \ . \quad (2.45)$$

▶**Exemple 2.15.** *Mélange de lois.* Si une variable aléatoire X suit une loi F_1 lorsqu'un certain événement A est réalisé (avec la probabilité p) et une loi F_2 dans le cas contraire, d'après (2.45) la fonction de répartition F de X est donnée par :

$$F(x) = pF_1(x) + (1-p)F_2(x) \, . \quad (2.46)$$

2.4 Exercices.

Exercice 2.1. Indiquer dans la liste suivante, quelles sont les fonctions susceptibles d'être des fonctions de répartition.

1. $\frac{1}{\pi} \arctan x + \frac{1}{2}$;
2. $\int_0^x e^{-t}\, dt$, $x \geq 0$;
3. $\int_0^x (4t - 2t^2 - 1)\, dt$, $0 \leq x \leq 2$;
4. $\int_0^x (4t - 2t^2)\, dt$, $0 \leq x \leq 2$

Exercice 2.2. *Durées de vie.* Soit T la durée de vie d'un composant. On a de bonnes raisons de croire que T est une variable aléatoire dont la densité de probabilité f suit la loi *exponentielle* :

$$f(t) = \begin{cases} Ae^{-t/\tau} & \text{si } t \geq 0 \, ; \\ 0 & \text{si } t < 0 \, . \end{cases}$$

- Calculer la constante A de façon à ce que $f(t)$ soit effectivement une densité de probabilité.
- Calculer la moyenne μ et l'écart type σ de T.
- On suppose que $\tau = 5\,\text{min}$, calculer la probabilité pour que le composant ait une durée de vie supérieure à 10 minutes.
- Calculer la probabilité : $\Pr\{T > 10\,\text{min} \,|T > 5\,\text{min}\}$, pour que le composant fonctionne plus de 10 minutes *sachant* qu'il a déjà fonctionné 5 minutes.

- Plus généralement, trouver la probabilité : $\Pr\{T > a+b | T > a\}$ pour que le composant fonctionne au moins un temps b de plus, sachant qu'il a déjà fonctionné pendant un temps a.

Exercice 2.3. *Téléphone.* Une compagnie téléphonique taxe ses clients par tranches de 3 minutes. Soit T le temps, exprimé en minutes, passé au téléphone par un client. On peut raisonnablement penser que la stratégie tarifaire de la compagnie induise un comportement chez le client tel que la fonction de répartition de T présente des discontinuités au voisinage des multiples de 3 minutes. Cette fonction de répartition pourrait ressembler à celle-ci :

$$F(t) = 1 - \tfrac{1}{2}e^{-t/3} - \tfrac{1}{2}e^{-\lfloor t/3 \rfloor}, \quad x \geq 0.$$

Dans cette expresion t est le temps de conversation exprimé en minutes et $\lfloor t \rfloor$ est la partie entière de t.

- Vérifier que F est bien une fonction de répartition et tracer son graphe.
- Quelle est la probabilité pour que la durée de conversation T soit : plus grande que 6 minutes ; plus petite que 4 minutes ; égale à 3 minutes ?
- Quelle est la probabilité condionnelle que la durée de conversation soit : plus petite que 9 minutes sachant qu'elle a déjà duré 5 minutes ; plus grande que 5 minutes sachant que de toutes façons elle sera interrompue si elle dépasse 9 minutes.

[voir Parzen [54] p.171]

Exercice 2.4. Donner un exemple de loi absolument continue, pourvue d'une densité et pour laquelle le mode n'est pas solution des équations (2.23).

Exercice 2.5. Généraliser les équations (2.29) en montrant que, dès que $k \geq 2$, les moments non-centrés μ'_k et les moments centrés μ_k s'expriment les uns en fonction des autres de la façon suivante :

$$\mu_k = \sum_{r=0}^{k-2} (-)^r C_k^r \mu^r \mu'_{k-r} - (-)^k (k-1)\mu^k, \quad \mu'_k = \sum_{r=0}^{k-2} C_k^r \mu^r \mu_{k-r} + \mu^k,$$

où C_k^r est le coefficient du binôme et μ la moyenne.

Exercice 2.6. *Un problème soumis à Pascal.* On renouvelle une expérience jusqu'à l'obtention d'un certain résultat. Si p désigne la probabilité d'obtenir le résultat cherché en un essai, trouver la moyenne et l'écart type du nombre N d'épreuves nécessaires pour obtenir le premier succès. On supposera que les épreuves sont indépendantes.

Exercice 2.7. Un client se présente devant un guichet pour être servi. La probabilité pour que le guichet soit libre est p, si le guichet est occupé le client attend. la loi suivie par le temps d'attente est une loi exponentielle de fonction de répartition $F(t) = 1 - e^{-\lambda t}$. Donner la fonction de répartition du temps d'attente T ainsi que la moyenne et l'écart type de cette variable aléatoire.

Chapitre 3

Variables aléatoires à plusieurs dimensions.

3.1 Un couple de variables aléatoires.

3.1.1 Fonction de répartition.

La fonction de répartition (bidimensionnelle ou 2D) F_{XY} d'un couple de variables aléatoires (X, Y) à valeurs dans $\Omega \subseteq \mathbb{R}^2$, est définie comme la probabilité pour que X ne dépasse pas le seuil x et que Y ne dépasse pas le seuil y :

$$F_{XY}(x, y) = \Pr\{X \leq x,\ Y \leq y\} \cdot \tag{3.1}$$

C'est la probabilité pour que le point aléatoire de coordonnées (X, Y) se trouve dans le quadrant $\mathcal{D}$ correspondant à l'événement $\{X \leq x, Y \leq y\}$, et représenté hachuré sur la figure 3.1. Comme d'habitude, en l'absence d'ambiguïté sur les variables aléatoires auxquelles la fonction F_{XY} se rapporte, on notera celle-ci simplement F. Notons que l'on a les relations suivantes :

$$F(-\infty, y) = F(x, -\infty) = 0\ , \quad \text{et} \quad F(\infty, \infty) = 1. \tag{3.2}$$

Nous appellerons quelquefois « *fonction de répartition conjointe* » la fonction F afin de la distinguer d'éventuelles autres fonctions de répartition.

3.1.2 Probabilité associée à un rectangle.

A l'aide de la fonction de répartition, nous pouvons calculer la probabilité pour que le couple (X, Y) se trouve à l'intérieur du rectangle défini par les relations $x_1 < X \leq x_2$ et $y_1 < Y \leq y_2$ (voir figure 3.2) :

$$\Pr\{x_1 < X \leq x_2, y_1 < Y \leq y_2\} = F(x_2, y_2) - F(x_1, y_2) - F(x_2, y_1) + F(x_1, y_1)\,. \tag{3.3}$$

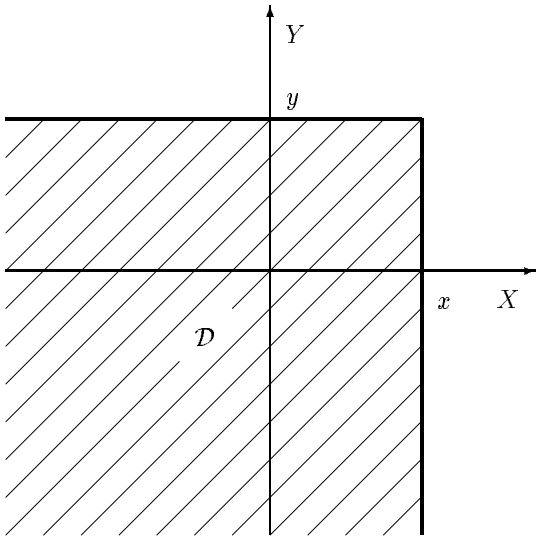


FIG. 3.1: *Domaine de définition de la fonction de répartition 2D.*

3.1.3 Densité de probabilité.

En posant $x_1 = x, y_1 = y, x_2 = x + \Delta x$ et $y_2 = y + \Delta y$ dans l'équation précédente, il vient au deuxième ordre en $\Delta x \Delta y$:

$$\Pr\{x < X \leq x + \Delta x, y < Y \leq y + \Delta y\} \simeq \frac{\partial^2}{\partial x \partial y} F(x, y) \Delta x \Delta y \,, \tag{3.4}$$

et en faisant tendre Δx et Δy vers 0 on obtient :

$$\lim_{\Delta x, \Delta y \to 0} \frac{\Pr\{x < X \leq x + \Delta x, y < Y \leq y + \Delta y\}}{\Delta x \Delta y} = \frac{\partial^2}{\partial x \partial y} F(x, y) \,. \tag{3.5}$$

La densité de probabilité du couple aléatoire (X, Y), ou comme il est courant de le dire, leur densité de probabilité conjointe, est définie (si elle existe) par :

$$f(x, y) = \frac{\partial^2}{\partial x \partial y} F(x, y) \,. \tag{3.6}$$

La fonction de répartition 2D se calcule alors à partir de la densité de probabilité à l'aide de l'expression suivante :

$$F(x, y) = \iint_{\mathcal{D}} f(u, v) du dv = \int_{-\infty}^{x} \int_{-\infty}^{y} f(u, v) du dv \,. \tag{3.7}$$

3.1.4 Lois marginales.

La loi suivie par un des membres du couple (X, Y) est appelée *loi marginale* de ce couple et la fonction de répartition du couple nous permet de calculer la fonction de répartition d'un terme de ce membre, par exemple X. Par définition,

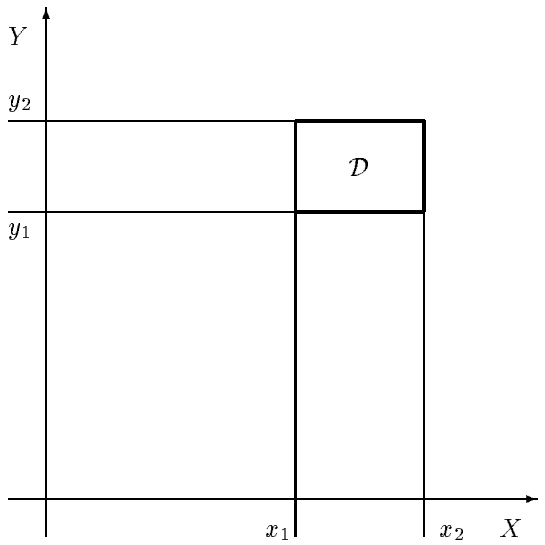


FIG. 3.2: *Probabilité p associée à un rectangle, $p = \iint_{\mathcal{D}} dF = F(x_2, y_2) - F(x_1, y_2) - F(x_2, y_1) + F(x_1, y_1)$.*

la fonction de répartition F_X de X est la probabilité pour que X ne dépasse pas le seuil x. En d'autres termes c'est la probabilité pour que le couple (X, Y) se trouve dans le demi-plan $\mathcal{D}_x$ représenté hachuré sur la figure 3.3. Il vient donc :

$$F_X(x) = F(x, \infty). \tag{3.8}$$

De la même façon on trouve la fonction de répartition $F_Y(y)$ de Y :

$$F_Y(y) = F(\infty, y). \tag{3.9}$$

Les fonctions $F_X(x)$ et $F_Y(y)$ calculées à partir d'une fonction de répartition bidimensionnelle $F(x, y)$ sont appelées *fonctions de répartition marginales de F.* Les densités de probabilité marginales f_X et f_Y sont par définition les dérivées des fonctions de répartition marginales :

$$f_X(x) = \frac{d}{dx} F_X(x) = \frac{d}{dx} F(x, \infty) = \frac{d}{dx} \int_{-\infty}^{x} du \int_{-\infty}^{\infty} f(u, v)\, dv\,, \tag{3.10a}$$

$$f_Y(x) = \frac{d}{dy} F_Y(y) = \frac{d}{dy} F(\infty, y) = \frac{d}{dy} \int_{-\infty}^{y} dv \int_{-\infty}^{\infty} f(u, v)\, du\,, \tag{3.10b}$$

et donc, sous réserve d'existence de f :

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)\, dy\ ; \qquad f_Y(y) = \int_{-\infty}^{\infty} f(x, y)\, dx\,. \tag{3.11}$$

3.1.5 Moments des lois 2D.

On introduit les moments du couple (X, Y) :

$$\mu'_{mn} = \int x^m y^n\, dF\,, \quad m, n \geq 0\,, \tag{3.12}$$

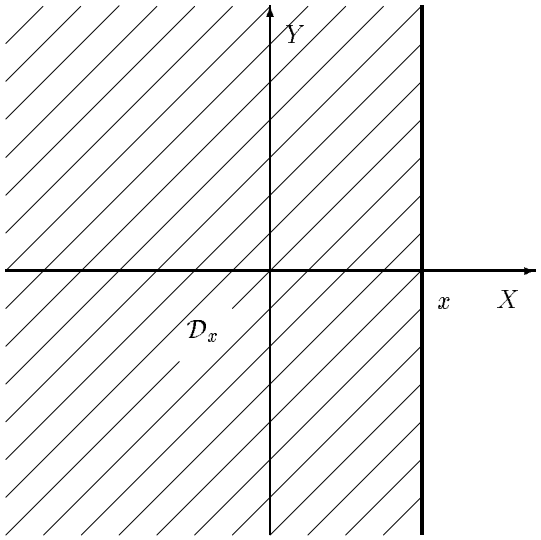


FIG. 3.3: *Domaine de définition de la fonction de répartition marginale* F_X.

et les moments centrés :

$$\begin{aligned} \mu_{mn} &= \int (x-\mu_X)^m (y-\mu_Y)^n \, dF\,, \quad m,n \geq 0, m+n>1; \\ \mu_{01} &= \mu'_{01}\,, \quad \mu_{10} = \mu'_{10}\,. \end{aligned} \tag{3.13}$$

Moyennes et variances. Le couple (μ'_{10}, μ'_{01}) ou identiquement le couple (μ_{10}, μ_{01}) des moments d'ordre 1 reçoit le nom de moyenne de la loi. On note $\boldsymbol{\mu}$ le vecteur colonne représentant cette moyenne. Les moments centrés μ_{20} et μ_{02} sont les variances de la loi, on les note habituellement σ_1^2 et σ_2^2, les quantités σ_1 et σ_2 sont les écart types.

Covariance et coefficient de corrélation. La quantité nouvelle μ_{11} reçoit le nom de « *covariance* » de (X,Y) et est notée $\mathrm{Cov}(X,Y)$. Il est immédiat d'établir la relation :

$$\mu_{11} = \mu'_{11} - \mu'_{10}\mu'_{01}\,. \tag{3.14}$$

Le coefficient de corrélation ρ est défini par :

$$\rho = \frac{\mu_{11}}{\sqrt{\mu_{20}\mu_{02}}}\,. \tag{3.15}$$

Le coefficient de corrélation possède les propriétés suivantes.

1. D'après l'inégalité de Cauchy-Schwarz, il est facile de voir que $|\rho| \leq 1$.
2. On montre que si $|\rho| = 1$, alors les variables aléatoires X et Y sont liées (presque sûrement) par une relation affine :

$$\frac{Y-\mu_Y}{\sqrt{\mathrm{Var}(Y)}} = \rho \frac{X-\mu_X}{\sqrt{\mathrm{Var}(X)}}\,. \tag{3.16}$$

Le coefficient de corrélation ρ, peut être considéré comme la mesure de la dépendance affine de X et Y, (on dit souvent par abus de langage, « dépendance linéaire », étant entendu que l'origine des axes a été ramenée sur la moyenne).

3. Si le coefficient de corrélation de (X, Y) est nul, on dit alors que ces variables aléatoires sont « *non-corrélées* ».

Matrice des variances-covariances.

On regroupe les moments centrés d'ordre 2 dans une matrice symétrique $\boldsymbol{V}$ appelée matrice des variances-covariances. Elle s'écrit de la façon suivante :

$$\boldsymbol{V} = \begin{pmatrix} \mu_{20} & \mu_{11} \\ \mu_{11} & \mu_{02} \end{pmatrix} \tag{3.17}$$

ou avec des notations plus habituelles :

$$\boldsymbol{V} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} . \tag{3.18}$$

Cette matrice est définie positive si $\sigma_1^2 \neq 0, \sigma_2^2 \neq 0$ et $|\rho| \neq 1$, c'est-à-dire si $\det \boldsymbol{V} \neq 0$.

3.1.6 Moments des lois marginales.

Les moments non-centrés et centrés des lois marginales sont égaux aux moments de la loi 2D ne faisant intervenir que la variable impliquée dans la loi marginale. Par exemple les moments centrés d'ordre i de X sont égaux aux μ_{i0} de la loi 2D et ceux de Y aux μ_{0i} de cette même loi. En particulier la loi marginale de X a pour moyenne $\mu_X = \mu_{10}$ et pour variance $\sigma_1^2 = \mu_{20}$ et celle de Y a pour moyenne $\mu_Y = \mu_{01}$ et pour variance $\sigma_2^2 = \mu_{02}$. Montrons-le pour μ_{10} :

$$\mu_{10} = \int_\Omega x f(x, y)\, dx dy = \int_{-\infty}^{\infty} x\, dx \int_{-\infty}^{\infty} f(x, y)\, dy ,$$

la dernière intégrale n'est autre que la loi marginale de X (voir l'équation (3.11)), d'où :

$$\mu_{10} = \int_{-\infty}^{\infty} x f_X(x)\, dx = \mu_X . \tag{3.19}$$

3.1.7 Variables aléatoires indépendantes.

Si l'on applique la règle des probabilité composées à l'événement $\{X \leq x, Y \leq y\}$, on obtient par exemple $\Pr\{X \leq x, Y \leq y\} = \Pr\{X \leq x\} \Pr\{Y \leq y | X \leq x\}$. Convenons de noter $F_{Y|X\leq x}(y, x)$ la dernière probabilité introduite dans l'expression précédente. Celle-ci s'écrit alors $F(x, y) = F_X(x) F_{Y|X\leq x}(y, x)$. Il est possible que la fonction $F_{Y|X\leq x}$ ne dépende pas de la valeur x. Elle représente alors la fonction de répartition F_Y de la variable Y. Cette remarque nous conduit à introduire la définition suivante.

Les variables aléatoires X et Y sont dites *indépendantes*, si leur fonction de répartition conjointe F peut être mise sous la forme du produit cartésien de leurs fonctions de répartition marginales soit $F = F_X \otimes F_Y$, c'est-à-dire si :

$$F(x,y) = F_X(x)F_Y(y) \ . \tag{3.20}$$

Si la fonction de répartition 2D admet une densité de probabilité alors, en appliquant la définition (3.6) à l'équation précédente, on montre que deux variables aléatoires sont indépendantes si on peut séparer leur densité de probabilité conjointe en un produit de deux densités, l'une ne dépendant que de x et l'autre que de y :

$$f(x,y) = f_X(x)f_Y(y) \ . \tag{3.21}$$

Ceci exprime que la densité 2D f est le produit cartésien des densités 1D, soit $f = f_X \otimes f_Y$.

L'indépendance implique la non corrélation, mais l'inverse n'est pas nécessairement vrai. Démontrons la première partie de cette affirmation. On a toujours $\mu_{11} = \mu'_{11} - \mu'_{10}\mu'_{01}$. Calculons μ'_{11} :

$$\mu'_{11} = \iint xyf(x,y)\,dxdy = \int xf_X(x)\,dx \int yf_Y(y)\,dy = \mu'_{10}\mu'_{01},$$

d'où $\mu_{11} = 0$ et donc $\rho = 0$ si $\mu_{10} \neq 0$ et $\mu_{01} \neq 0$.

▶**Exemple 3.1.** *Loi normale 2D.* Supposons que le couple de variables aléatoires (X,Y) suive une loi de densité de probabilité donnée par l'expression :

$$f(x,y) = \frac{1}{2\pi\sigma_1\sigma_2} \exp\left\{-\left(\frac{(x-\mu_1)^2}{2\sigma_1^2} + \frac{(y-\mu_2)^2}{2\sigma_2^2}\right)\right\}. \tag{3.22}$$

Cette densité de probabilité peut être mise sous la forme d'un produit de deux densités de probabilité portant séparément sur chacune des variables :

$$f(x,y) = \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left\{-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right\} \frac{1}{\sqrt{2\pi}\sigma_2} \exp\left\{-\frac{(x-\mu_2)^2}{2\sigma_2^2}\right\}. \tag{3.23}$$

Ceci montre que les variables aléatoires X et Y sont indépendantes. Lorsque cette opération est possible, on doit prendre garde à séparer la densité conjointe en un produit de deux fonctions qui sont bien des densités de probabilité, c'est-à-dire dont l'intégrale sur $\mathbb{R}$ soit égale à 1. C'est ce qui a été fait ici en séparant convenablement en deux le terme $1/2\pi\sigma_1\sigma_2$.

3.1.8 Lois conditionnelles associées à une loi 2D.

La fonction de répartition du couple aléatoire (X,Y) permet de calculer les probabilités associées à ce couple lorsqu'il est astreint à n'occuper qu'un sous-ensemble Ω' du domaine de définition original Ω. La fonction de répartition du couple (X,Y) restreint au domaine Ω' est notée $F_{\Omega'}(x,y|\Omega')$. Elle est égale à la probabilité conditionnelle $\Pr\{X \leq x, Y \leq y | (X,Y) \in \Omega'\}$. Comme dans le cas 1D, on calcule cette dernière valeur grâce à la règle des probabilités composées $\Pr\{A|B\} = \Pr\{AB\} \,/\, \Pr\{B\}$, où ici l'événement A est $(X,Y) \in \mathcal{D}$ et l'événement B est $(X,Y) \in \Omega'$. On obtient alors :

$$F_{\Omega'}(x,y|\Omega') = \Pr\{X \leq x, Y \leq y | (X,Y) \in \Omega'\} = \frac{\Pr\{X \leq x, Y \leq y, (X,Y) \in \Omega'\}}{\Pr\{(X,Y) \in \Omega'\}} \ . \tag{3.24}$$

Cette quantité se calcule facilement à l'aide de la fonction de répartition du couple (X, Y) et l'on obtient :

$$F_{\Omega'}(x, y|\Omega') = \frac{\iint_{\mathcal{D}\cap\Omega'} dF(u, v)}{\iint_{\Omega'} dF(u, v)} \cdot \tag{3.25}$$

En introduisant la fonction indicatrice $\mathbf{1}_{\Omega'}$ qui vaut 1 sur le domaine Ω' et 0 ailleurs, on obtient :

$$F_{\Omega'}(x, y|\Omega') = \frac{\iint_{\mathcal{D}} \mathbf{1}_{\Omega'}(u, v)\, dF(u, v)}{\iint_{\Omega} \mathbf{1}_{\Omega'}(u, v)\, dF(u, v)} \cdot \tag{3.26}$$

Le point important à retenir dans cette façon de présenter les choses est que les domaines d'intégration ont été ramenés à des domaines connus, à savoir le domaine $\mathcal{D}$ illustré sur la figure 3.1 et le domaine de définition Ω. Rappelons que le domaine $\mathcal{D}$ dépend de x et de y, et que c'est par cet intermédiaire que $F_{\Omega'}$ dépend de x et de y.

Le cas dégénéré.

Envisageons maintenant le cas où le support de Ω' est de mesure nulle dans $\mathbb{R}^2$, par exemple quand Ω' dégénère vers une relation quelconque liant X à Y. Soit g cette relation. On a $g(X, Y) = 0$. La fonction $\mathbf{1}_{\Omega'}(x, y)/\iint_{\Omega} \mathbf{1}_{\Omega'}$ tend vers la distribution de Dirac $\delta(g(x, y))$ et il vient :

$$F_{\Omega'}(x, y|\Omega') = \frac{\iint_{\mathcal{D}} \delta(g(u, v))\, dF(u, v)}{\iint_{\Omega} \delta(g(u, v))\, dF(u, v)} \cdot \tag{3.27}$$

Si la loi admet une densité on a :

$$F_{\Omega'}(x, y|\Omega') = \frac{\iint_{\mathcal{D}} f(u, v)\delta(g(u, v))\, du dv}{\iint_{\Omega} f(u, v)\delta(g(u, v))\, du dv} \cdot \tag{3.28}$$

Aux fonctions de répartition ainsi définies, on peut, sous réserve d'existence, associer les densités de probabilité :

$$f_{\Omega'}(x, y|\Omega') = \frac{\partial^2}{\partial x \partial y} F(x, y|\Omega') \cdot \tag{3.29}$$

Si la loi 2D admet une densité de probabilité f, on obtient après dérivation de (3.28) :

$$f_{\Omega'}(x, y|\Omega') = \frac{f(x, y)\delta(g(x, y))}{\iint_{\Omega} f(u, v)\delta(g(u, v))\, du dv} \cdot \tag{3.30}$$

3.1.9 Lois conditionnelles d'une coupe.

On réserve, souvent, l'appellation de fonction de répartition conditionnelle ou de densité de probabilité conditionnelle au cas où le support de $g(X, Y)$ se réduit à l'équation $Y = Cste$ ou $X = Cste$, c'est-à-dire au cas où l'on considère une coupe parallèle aux axes, faite à travers la densité de probabilité bidimensionnelle. On s'intéresse alors aux lois de Y sachant que X vaut une certaine

valeur x_0, ou bien aux lois de X sachant que Y vaut y_0. Etudions ce dernier cas ; la relation $g(X,Y)=0$ s'écrit alors $Y-y_0=0$ et en introduisant cette relation dans l'équation (3.28) on trouve la fonction de répartition conditionnelle :

$$F(x,y|Y=y_0)=\frac{\iint_{\mathcal{D}} f(u,v)\delta(v-y_0)\,dudv}{\iint_{\Omega} f(u,v)\delta(v-y_0)\,dudv}\,. \tag{3.31}$$

Pour le numérateur, l'intégration sur v donne 0 si y est inférieur à y_0, et $f(u,y_0)$ dans le cas contraire. Pour le dénominateur, cette intégration donne toujours $f(u,y_0)$. Il vient donc :

$$F(x,y|Y=y_0)=\frac{\displaystyle\int_{-\infty}^{x} f(u,y_0)\mathbf{1}_{[y_0,\infty[}(u,y)\,du}{\displaystyle\int_{-\infty}^{\infty} f(u,y_0)\,du}\,. \tag{3.32}$$

La fonction indicatrice du numérateur est la fonction H de Heaviside. Elle ne dépend pas de u et peut donc être sortie de l'intégrale, d'où :

$$F(x,y|Y=y_0)=\frac{\displaystyle\int_{-\infty}^{x} f(u,y_0)\,du}{\displaystyle\int_{-\infty}^{\infty} f(u,y_0)\,du}H(y-y_0). \tag{3.33}$$

La fonction de répartition conditionnelle bidimensionnelle a été mise sous la forme d'un produit de deux fonctions de répartition à une dimension $F_{X|Y}(x|Y=y_0)$ et $F_Y(y)=H(y-y_0)$. La première est la fonction de répartition conditionnelle de X, la deuxième est la fonction de répartition marginale de Y, cette dernière rendant compte du fait que Y a été fixé à la valeur y_0. On a donc :

$$F_{X|Y}(x|Y=y_0)=\frac{\int_{-\infty}^{x} f(u,y_0)\,du}{\int_{-\infty}^{\infty} f(u,y_0)\,du}\,, \tag{3.34}$$

le dénominateur n'est autre que la densité marginale de Y en y_0, d'où :

$$F_{X|Y}(x|Y=y_0)=\frac{1}{f_Y(y_0)}\int_{-\infty}^{x} f(u,y_0)\,du \tag{3.35}$$

et la relation similaire pour la densité conditionnelle de la coupe $X=x_0$:

$$F_{Y|X}(y|X=x_0)=\frac{1}{f_X(x_0)}\int_{-\infty}^{y} f(x_0,v)\,dv\,. \tag{3.36}$$

Densité de probabilité conditionnelle d'une coupe.

On obtient les densités de probabilité conditionnelles par dérivation des fonctions de répartition correspondantes. Ainsi :

$$f_{X|Y}(x|Y=y_0)=\frac{\partial}{\partial x}F_{X|Y}(x|Y=y_0)\,, \tag{3.37}$$

et de la même façon :

$$f_{Y|X}(y|X=x_0)=\frac{\partial}{\partial y}F_{Y|X}(x|X=x_0)\,. \tag{3.38}$$

On trouve alors :

$$f_{X|Y}(x|Y=y_0) = \frac{f(x,y_0)}{\int_{-\infty}^{\infty} f(x,y_0)\,dx}\,, \tag{3.39a}$$

$$f_{Y|X}(y|X=x_0) = \frac{f(x_0,y)}{\int_{-\infty}^{\infty} f(x_0,y)\,dy}\,. \tag{3.39b}$$

En termes imagés, la densité conditionnelle (pour la condition $Y = y_0$ par exemple), est une coupe effectuée à travers la densité 2D pour la valeur y_0. La surface sous la coupe n'est en général pas égale à 1, il faut alors normaliser la coupe en la divisant par cette surface. C'est ce qu'expriment les équations (3.39a) et (3.39b).

S'il n'y a pas à craindre de confusion, on notera simplement $f_{X|Y}(x|y)$ et $f_{Y|X}(y|x)$ les densités conditionnelles des coupes $Y = y$ et $X = x$, étant sous-entendu que les valeurs y et x sont particulières. Avec cette notation on a :

$$f_{X|Y}(x|y) = \frac{f(x,y)}{f_Y(y)}\,, \qquad f_{Y|X}(y|x) = \frac{f(x,y)}{f_X(x)}\,. \tag{3.40}$$

Ces formules lient entre elles les densités conjointes, marginales et conditionnelles. Elles montrent en particulier que si les variables aléatoires X et Y sont indépendantes, alors les densités de probabilité conditionnelles sont égales aux densités de probabilité marginales. En effet l'indépendance implique $f(x,y) = f_X(x)f_Y(y)$ et il vient :

$$f_{X|Y}(x|y) = f_X(x), \quad f_{Y|X}(y|x) = f_Y(y)\,. \tag{3.41}$$

Ainsi, pour des variables aléatoires indépendantes, la connaissance de la valeur prise par une des variables aléatoires ne modifie pas la répartition de l'autre. Ce qui correspond bien à la notion intuitive d'indépendance.

Théorème de Bayes et formule des probabilités totales.

En éliminant $f(x,y)$ des équations (3.40) on trouve :

$$f_{X|Y}(x|y) = \frac{f_{Y|X}(y|x)f_X(x)}{f_Y(y)}\,, \tag{3.42}$$

ce qui est la traduction de la formule de Bayes, dans le langage des densités de probabilité. En remplaçant l'une ou l'autre des densités marginales par les expressions (3.11) et en utilisant (3.40), on obtient les formules dites des « *probabilités totales* » :

$$f_{X|Y}(x|y) = \frac{f_{Y|X}(y|x)f_X(x)}{\int_{-\infty}^{\infty} f_{Y|X}(y|x)f_X(x)\,dx}, \qquad f_{Y|X}(y|x) = \frac{f_{X|Y}(x|y)f_Y(y)}{\int_{-\infty}^{\infty} f_{X|Y}(x|y)f_Y(y)\,dy}\,. \tag{3.43}$$

Lois *a priori* et lois *a posteriori*.

Lorsqu'une variable aléatoire X suit une loi qui dépend de la détermination préalable d'une autre variable aléatoire Y, alors on appelle loi *a priori* la loi marginale $f_X(x)$ de X et loi *a posteriori* la loi conditionnelle $f_{X|Y}(x|y)$ de X.

La formule des probabilités totales dans son expression (3.43) lie entre elles les densités *a priori* et *a posteriori*.

On utilise la formule des probabilités totales dans le cas, très général, où la variable aléatoire Y est une grandeur physique non-directement observable mais déterminant la loi suivie par une grandeur observable X. La variable aléatoire Y caractérise en quelque sorte « l'état de la nature. » La loi *a priori* $f_Y(y)$ est une mesure de notre connaissance sur Y avant toute observation de X, et la loi *a posteriori* $f_{Y|X}(y|x)$ rend compte de la connaissance gagnée sur Y après l'observation $X = x$. Afin d'accroître notre connaissance sur Y à partir d'une observation de X, il faut naturellement connaître la loi suivie par X quand Y est connu, c'est-à-dire $f_{X|Y}(x|y)$. On peut alors calculer $f_{Y|X}$ à l'aide de la formule des probabilités totales. Cette fonction est une représentation du « crédit » que l'on peut accorder à l'hypothèse suivant laquelle la nature était dans l'état y préalablement à l'observation x.

Moyennes conditionnelles et courbes de régression.

On trouve l'expression de la moyenne et de la variance conditionnelle d'une coupe en appliquant aux lois conditionnelles les équations (2.25) et (2.28) définissant les moments. Il vient :

$$\eta_x = \int_{-\infty}^{\infty} y f_Y(y|x)\, dy\,, \tag{3.44}$$

$$\sigma_x^2 = \int_{-\infty}^{\infty} (y - \eta_x)^2 f_Y(y|x)\, dy\,, \tag{3.45}$$

et des expressions analogues pour η_y et σ_y^2. Les moyennes conditionnelles η_x et η_y sont des fonctions $\phi_Y(x)$ et $\phi_X(y)$, appelées la première « courbe de régression de Y par rapport à X », et la deuxième « courbe de régression de X par rapport à Y ».

3.2 Plusieurs variables aléatoires.

Pour un nombre fini de variables aléatoires, $(X_1, \ldots, X_n)$, on généralise sans peine les notions introduites pour un couple de variables aléatoires. Un tel ensemble de variables aléatoires peut aussi être considéré comme les n composantes d'un vecteur aléatoire à n dimensions $\boldsymbol{X} \in \mathbb{R}^n$.

3.2.1 Vecteurs aléatoires et notation matricielle.

On notera souvent un ensemble de variables aléatoires $(X_1, \ldots, X_n)$ par le même symbole en caractère gras privé d'indice, dans notre cas $\boldsymbol{X}$. Nous dirons que $\boldsymbol{X}$ est un vecteur aléatoire, étant entendu que $\boldsymbol{X}$ représente les coordonnées du vecteur mis sous la forme d'une colonne. Le symbole $\boldsymbol{X}^t$ désigne le même vecteur, mais ses coordonnées étant mises sous forme d'une ligne.

3.2.2 Fonction de répartition.

La fonction de répartition $F_{X_1 X_2 \cdots X_n}$ est définie par l'expression :

$$F_{X_1 X_2 \cdots X_n}(x_1, x_2, \ldots, x_n) = \Pr\{X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n\}\,, \tag{3.46}$$

et possède les propriétés suivantes (nous la notons simplement F pour plus de commodité)

Propriétés de la fonction de répartition.

1. La fonction F est définie sur $\mathbb{R}^n$ et envoie $\mathbb{R}^n$ sur $[0,1]$, $F : \mathbb{R}^n \to [0,1]$.
2. $F(x_1, \ldots, x_n)$ est une fonction non-décroissante de chacune des variables.
3. $F(x_1, \ldots, x_n)$ est continue à droite en chacune des variables.
4. $F(x_1, \ldots, x_n) = 0$ si au moins une des variables vaut $-\infty$.
5. $F(x_1, \ldots, x_n) = 1$ si toutes les variables valent $+\infty$.

3.2.3 Probabilité d'un hyper-rectangle.

La probabilité pour que les n variables aléatoires X_k soient comprises dans les intervalles $\forall k; a_k < X_k \leq b_k$, c'est-à-dire pour que le vecteur aléatoire $\boldsymbol{X}$ « tombe » dans l'hyper-rectangle $\forall k; a_k < x_k \leq b_k$, est donnée par la formule :

$$\Pr\{\forall k; a_k < X_k \leq b_k\} = \sum_{\epsilon_1=0}^{1} \sum_{\epsilon_2=0}^{1} \cdots \sum_{\epsilon_n=0}^{1} (-1)^{\sum_{k=1}^{n} \epsilon_k} F(c_1, \ldots, c_n) , \quad (3.47)$$

où $c_k = \epsilon_k a_k + (1-\epsilon_k) b_k$. Les c_k sont les coordonnées des sommets de l'hyper-rectangle et la somme s'étend sur les 2^n sommets de cet hyper-rectangle. En posant $\forall k;\ b_k = a_k + h_k$, on peut exprimer (3.47) à l'aide de l'opérateur aux différences finies Δ_h^k portant sur la k^{e} variable de F et tel que :

$$\Delta_h^k F(\ldots, x_k, \ldots) = F(\ldots, x_k + h, \ldots) - F(\ldots, x_k, \ldots) . \quad (3.48)$$

Il vient alors :

$$\Pr\{\forall k; a_k < X_k \leq a_k + h_k\} = \Delta_{h_1}^1 \Delta_{h_2}^2 \cdots \Delta_{h_n}^n F(a_1, \ldots, a_n) . \quad (3.49)$$

3.2.4 Densité de probabilité.

Quand elle existe, la densité de probabilité est telle que :

$$F_{X_1 X_2 \cdots X_n}(x_1, x_2, \ldots, x_n) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} \cdots \int_{-\infty}^{x_n} f_{X_1 X_2 \cdots X_n}(u_1, u_2, \ldots, u_n)\, du_1 du_2 \cdots du_n . \quad (3.50)$$

Ce qui permet de calculer $f_{X_1 X_2 \cdots X_n}$ connaissant $F_{X_1 X_2 \cdots X_n}$:

$$f_{X_1 X_2 \cdots X_n}(x_1, x_2, \ldots, x_n) = \frac{\partial^n}{\partial x_1 \partial x_2 \ldots \partial x_n} F_{X_1 X_2 \cdots X_n}(x_1, x_2, \ldots, x_n) . \quad (3.51)$$

La densité de probabilité possède les propriétés suivantes :

1. $f_{X_1 X_2 \cdots X_n}(x_1, x_2, \ldots, x_n) \geq 0$ pour presque tous les $x_1, x_2, \ldots, x_n$.
2. $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{X_1 X_2 \cdots X_n}(x_1, x_2, \ldots, x_n)\, dx_1 dx_2 \cdots dx_n = 1$.

Pour plus de commodité nous noterons, la plupart du temps, F la fonction de répartition, et f la densité de probabilité.

3.2.5 Lois marginales.

Les fonctions de répartition marginales d'un ensemble de $k < n$ variables aléatoires s'obtiennent par passage à la limite ∞ des $n-k$ autres variables. Par exemple la fonction de répartition de la variable aléatoire X_1 s'obtient comme fonction de répartition marginale de F à l'aide de la formule suivante :

$$F_{X_1}(x) = F(x, \infty, \ldots, \infty) \ . \tag{3.52}$$

Les densités de probabilité marginales des k variables aléatoires s'expriment à l'aide de la densité $f_{X_1 X_2 \cdots X_n}$, quand elle existe, par intégration sur les $n-k$ variables aléatoires restantes. Par exemple, pour la variable aléatoire X_k :

$$f_{X_k}(x) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{X_1 X_2 \cdots X_n}(x_1, \ldots, x_{k-1}, x, x_{k+1}, \ldots, x_n) \\ dx_1 \cdots dx_{k-1} dx_{k+1} \cdots dx_n \ . \tag{3.53}$$

3.2.6 Moments.

Les moments non-centrés sont définis par la formule :

$$\mu'_{n_1 n_2 \cdots n_n} = \int_{-\infty}^{\infty} x_1^{n_1} x_2^{n_2} \cdots x_n^{n_n} f_{X_1 X_2 \cdots X_n}(x_1, x_2, \ldots, x_n)\, dx_1 dx_2 \cdots dx_n \ . \tag{3.54}$$

La somme des exposants $\sum_i n_i$ est l'ordre du moment. Dans le cas où certains indices n_i sont nuls, les moments ainsi calculés sont les moments des variables aléatoires correspondant aux indices non nuls. Montrons cela pour le moment $\mu'_{10\cdots0}$. D'après ce que nous venons de dire, ce doit être la moyenne μ_1 de la variable aléatoire X_1. En effet :

$$\mu'_{10\cdots0} = \int_{-\infty}^{\infty} x_1 f_{X_1 X_2 \cdots X_n}(x_1, x_2, \ldots, x_n)\, dx_1 dx_2 \cdots dx_n \ ,$$

mais l'intégration sur les variables $x_2, \ldots, x_n$ n'est autre, d'après l'équation (3.53), que la densité marginale de f pour la variable aléatoire X_1. Il reste donc :

$$\mu'_{10\cdots0} = \int_{-\infty}^{\infty} x_1 f_{X_1}(x_1)\, dx_1 \ .$$

Ce qui montre que $\mu'_{10\cdots0}$ correspond bien à la moyenne μ_{X_1} de X_1. On introduit maintenant les moments centrés qui sont par définition :

$$\mu_{n_1 n_2 \cdots n_n} = \int_{-\infty}^{\infty} (x_1 - \mu_{X_1})^{n_1} (x_2 - \mu_{X_2})^{n_2} \cdots (x_n - \mu_{X_n})^{n_n} \\ f_{X_1 X_2 \cdots X_n}(x_1, x_2, \ldots, x_n)\, dx_1 dx_2 \cdots dx_n \ . \tag{3.55}$$

On démontrerait de la même manière que, par exemple, le moment centré $\mu_{20\cdots0}$ correspond à la variance de la variable aléatoire X_1. La densité de probabilité de X_1 étant calculée comme densité marginale de $f_{X_1 X_2 \cdots X_n}$.

3.2.7 Matrice des variances-covariances.

Les moments d'ordre deux sont regroupés dans la matrice des *variances-covariances* $\boldsymbol{V}$, d'éléments :

$$v_{ij} = \int_{\Omega} (x_i - \mu_{X_i})(x_j - \mu_{X_j}) f(x_1, \ldots, x_n)\, dx_1 \cdots dx_n \tag{3.56}$$

Les éléments diagonaux de cette matrice sont les variances des X_i et les éléments non-diagonaux v_{ij} $i \neq j$ sont les covariances des couples (X_i, X_j). Toujours d'après l'inégalité de Cauchy-Schwarz on a : $v_{ij}^2 \leq v_{ii} v_{jj}$. Cela montre que la matrice des variances-covariances est définie non-négative. Les coefficients de corrélation ρ_{ij} des couples (X_i, X_j) sont définis par :

$$\rho_{ij} = \frac{v_{ij}}{\sqrt{v_{ii} v_{jj}}} \cdot \tag{3.57}$$

On pose habituellement $v_{ii} = \sigma_i^2$ ($\sigma_i \geq 0$), ce qui fait que la covariance s'écrit plutôt $v_{ij} = \rho_{ij} \sigma_i \sigma_j$.

Les moments des lois marginales de F étant égaux aux moments de F dont les seuls indices non-nuls sont ceux correspondants aux variables aléatoires sur lesquelles portent les lois marginales, la matrice des variances-covariances de ces lois marginales est donc une sous-matrice extraite de la matrice des variances-covariances de F en supprimant les lignes et les colonnes ne correspondant pas aux variables des lois marginales. Si, par exemple, le triplet aléatoire (X_1, X_2, X_3) a pour matrice des variances-covariances :

$$\boldsymbol{V} = \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \rho_{13}\sigma_1\sigma_3 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 & \rho_{23}\sigma_2\sigma_3 \\ \rho_{13}\sigma_1\sigma_3 & \rho_{23}\sigma_2\sigma_3 & \sigma_3^2 \end{pmatrix},$$

la matrice des variances-covariances du couple (X_1, X_3) s'obtient en supprimant la deuxième ligne et la deuxième colonne de $\boldsymbol{V}$:

$$\boldsymbol{V}_{13} = \begin{pmatrix} \sigma_1^2 & \rho_{13}\sigma_1\sigma_3 \\ \rho_{13}\sigma_1\sigma_3 & \sigma_3^2 \end{pmatrix}.$$

Cette remarque s'applique naturellement aussi au vecteur colonne représentant la moyenne.

3.2.8 Lois conditionnelles.

Le mécanisme qui nous a permis de trouver les lois conditionnelles dans le cas 2D s'applique également ici, nous ne le répéterons pas. Il en résulte, vis-à-vis d'une condition Ω', une formule analogue à l'équation (3.26). Nous ne considérerons plus en détail ici que le cas dégénéré correspondant à une coupe.

3.2.9 Lois conditionnelles des coupes.

Comme dans le cas 2D, une coupe correspond à la condition où certaines variables aléatoires d'un vecteur $(X_1, \ldots, X_{m+n})$ sont fixées, alors que les autres restent libres. Nous allons supposer, afin d'alléger l'écriture, que les variables libres correspondent aux indices les plus faibles de la liste des X_i. Le symbole

$\boldsymbol{X}_0$ représentera l'ensemble des m variables aléatoires libres et le symbole $\boldsymbol{X}_1$ représentera celui des n variables fixées. Notons de plus $\boldsymbol{y}$ l'ensemble $(y_1, \ldots, y_n)$ des valeurs prises par $\boldsymbol{X}_1$ et par $\boldsymbol{x}$ l'ensemble $(x_1, \ldots, x_m)$ des valeurs possibles de $\boldsymbol{X}_0$. Il vient :

$$F_{\boldsymbol{X}_0\boldsymbol{X}_1}(\boldsymbol{x}, \boldsymbol{y}) = F_{\boldsymbol{X}_1}(\boldsymbol{y}) F_{\boldsymbol{X}_0|\boldsymbol{X}_1}(\boldsymbol{x}|\boldsymbol{X}_1 = \boldsymbol{y}) , \tag{3.58}$$

où $F_{\boldsymbol{X}_0\boldsymbol{X}_1}$ désigne la loi conjointe des variables X_i et $F_{\boldsymbol{X}_1}$ désigne la loi marginale des variables fixées. Cette formule nous permet de calculer la loi conditionnelle. Si la loi marginale possède une densité, cette formule s'écrit :

$$F_{\boldsymbol{X}_0|\boldsymbol{X}_1}(\boldsymbol{x}|\boldsymbol{X}_1 = \boldsymbol{y}) = \frac{1}{f_{\boldsymbol{X}_1}(\boldsymbol{y})} \frac{\partial^n}{\partial^n \boldsymbol{y}} F_{\boldsymbol{X}_0\boldsymbol{X}_1}(\boldsymbol{x}, \boldsymbol{y}) . \tag{3.59}$$

La notation $\partial^n / \partial^n \boldsymbol{y}$ désigne la dérivation par rapport à toutes les variables fixées, c'est-à-dire $\partial^n / \partial y_1, \ldots, \partial y_n$. Si la densité conjointe des variables X_i existe, on trouve, en dérivant (3.59) par rapport aux x_i, la formule liant entre elles les densités conjointes, marginales et conditionnelles :

$$f_{\boldsymbol{X}_0|\boldsymbol{X}_1}(\boldsymbol{x}|\boldsymbol{X}_1 = \boldsymbol{y}) = \frac{1}{f_{\boldsymbol{X}_1}(\boldsymbol{y})} f_{\boldsymbol{X}_0\boldsymbol{X}_1}(\boldsymbol{x}, \boldsymbol{y}) . \tag{3.60}$$

3.2.10 Variables aléatoires indépendantes.

Les variables aléatoires $X_1, \ldots, X_n$ seront dites « *mutuellement indépendantes* » ou « *indépendantes dans leur ensemble* » ou plus simplement « *indépendantes* », si la fonction de répartition conjointe peut être mise sous la forme du produit des fonction de répartition marginales :

$$F(x_1, x_2, \ldots, x_n) = F_{X_1}(x_1) F_{X_2}(x_2) \cdots F_{X_n}(x_n) . \tag{3.61}$$

- Si les variables aléatoires $X_1, X_2, \ldots, X_n$ sont indépendantes, alors elles sont deux à deux indépendantes, c'est-à-dire :

$$F_{X_i X_j}(x_i, x_j) = F_{X_i}(x_i) F_{X_j}(x_j), \quad \forall i, j; i \neq j . \tag{3.62}$$

  Démontrons cette propriété pour $i = 1, j = 2$. Par définition des fonctions de répartition marginales on a :

$$F_{X_1 X_2}(x_1, x_2) = F(x_1, x_2, \infty, \ldots, \infty).$$

  D'après (3.61) cette expression est égale à :

$$F_{X_1}(x_1) F_{X_2}(x_2) F_{X_3}(\infty) \cdots F_{X_n}(\infty);$$

  mais $F_{X_i}(\infty) = 1$, d'où on tire (3.62).

- Si les variables aléatoires $X_1, \ldots, X_n$ sont deux à deux indépendantes, elles ne sont pas nécessairement indépendantes dans leur ensemble. En revanche, les coefficients de corrélation ρ_{ij} des couples $(X_i, X_j), i \neq j$ sont nuls et la matrice des variances-covariances est diagonale. On dit alors que les variables aléatoires sont mutuellement non-corrélées. Des variables aléatoires mutuellement non-corrélées ne sont pas nécessairement deux à deux indépendantes.

3.3 Plusieurs vecteurs aléatoires.

Les notions introduites pour les vecteurs aléatoires se généralisant sans peine à plusieurs vecteurs aléatoires, nous n'introduirons ici que la notion nouvelle de matrice de covariance.

3.3.1 La matrice de covariance.

Soient deux vecteurs aléatoires $\boldsymbol{X}$ et $\boldsymbol{Y}$. On définit leur *matrice de covariance* $\boldsymbol{C}_{XY}$ dont les éléments c_{ij} sont donnés par l'expression :

$$c_{ij} = \int (x_i - \mu_{X_i})(y_j - \mu_{Y_j})\, dF\,, \tag{3.63}$$

l'intégrale étant calculée à l'aide de la loi F du couple $(\boldsymbol{X}, \boldsymbol{Y})$.

Chapitre 4

Changement de variable aléatoire.

4.1 Une variable et une fonction.

Soit $X(\omega)$ une variable aléatoire associée à l'événement élémentaire ω. A ce même événement élémentaire ω, on fait correspondre une autre variable aléatoire $Y(\omega)$ telle que $Y = \varphi(X)$. La fonction φ ainsi introduite, définit ce que l'on appelle un changement de variable aléatoire. Le problème que nous nous posons maintenant est de déterminer la fonction de répartition G de Y et éventuellement sa densité de probabilité g, connaissant la fonction de répartition F de X.

Par définition la fonction de répartition $G(y)$ de Y est la probabilité pour que Y ne dépasse pas le seuil y. On a :

$$G(y) \equiv \Pr\{Y(\omega) \leq y\} = \Pr\{\varphi(X(\omega)) \leq y\} \ , \tag{4.1}$$

notre problème sera résolu dès que l'on saura trouver les solutions de l'inégalité $\varphi(X(\omega)) \leq y$. Nous allons donner plusieurs exemples de telles solutions. Afin d'alléger l'exposé, nous noterons à partir de maintenant simplement X la variable aléatoire $X(\omega)$. Dans le même but, lorsque nous parlerons d'une densité de probabilité, nous supposerons implicitement qu'elle existe.

4.1.1 Variables aléatoires continues.

Cas où φ est univoque, dérivable et croissante.

Dans ce cas la fonction réciproque φ^{-1} existe et la solution de l'inégalité $\varphi(X) \leq y$ est $X \leq \varphi^{-1}(y)$. Il vient alors d'après (4.1) :

$$G(y) = \Pr\left\{X \leq \varphi^{-1}(y)\right\} = F(\varphi^{-1}(y)) \ . \tag{4.2}$$

On obtient la densité de probabilité en dérivant G par rapport à y, et en posant $x = \varphi^{-1}(y)$ la solution unique de l'équation $y = \varphi(x)$, il vient :

$$g(y) = \frac{dG}{dy} = \frac{dF}{dx}\frac{d\varphi^{-1}}{dy} \ . \tag{4.3}$$

La première dérivée est par définition la densité de probabilité f de X et la seconde l'inverse de la dérivée de φ au point $x = \varphi^{-1}(y)$, de sorte que l'on obtient g par la formule :

$$g(y) = \frac{f(x)}{\left|\frac{d\varphi}{dx}\right|_{x=\varphi^{-1}(y)}} \tag{4.4}$$

Cas où φ est univoque, dérivable et décroissante.

Dans ce cas la solution de $\varphi(X) \leq y$ est $X \geq \varphi^{-1}(y)$ et il vient :

$$G(y) = \Pr\left\{X \geq \varphi^{-1}(y)\right\} = 1 - \Pr\left\{X < \varphi^{-1}(y)\right\} .$$

L'événement $X = \varphi^{-1}(y)$ étant de mesure nulle on trouve alors l'expression de la fonction de répartition de Y :

$$G(y) = 1 - \Pr\left\{X \leq \varphi^{-1}(y)\right\} = 1 - F(\varphi^{-1}(y)) , \tag{4.5}$$

et de sa densité de probabilité :

$$g(y) = \frac{-f(x)}{\left|\frac{d\varphi}{dx}\right|_{x=\varphi^{-1}(y)}} \tag{4.6}$$

Cas général des changements de variable bijectifs.

Dans le cas général des changements de variables effectués à l'aide d'une fonction φ univoque, on a en rassemblant en une seule les deux formules précédentes :

$$\boxed{g(y) = f(x)\frac{1}{\left|\frac{d\varphi}{dx}\right|_{x=\varphi^{-1}(y)}}} \tag{4.7}$$

La table 4.1 résume les principaux résultats de ce paragraphe.

▶**Exemple 4.1.** *Changement de variable linéaire $Y = aX$.* Dans ce cas la fonction φ est telle que $y = \varphi(x) = ax$ d'où $\varphi^{-1}(y) = y/a$. En appliquant la formule (4.7) précédente on obtient alors :

$$\frac{d\varphi}{dx} = a \quad \text{et donc} \quad g(y) = f(x)\frac{1}{|a|} \quad \text{d'où} \quad g(y) = \frac{1}{|a|} f\left(\frac{y}{a}\right) . \tag{4.8}$$

Cas des fonctions non-univoques.

Si la fonction φ n'est pas univoque mais reste dérivable, les valeurs de X satisfaisant l'inégalité $\varphi(X) \leq y$ se présentent sous la forme d'intervalles disjoints $[a_k, b_k]$, en nombre éventuellement infini mais dénombrable. On a $\varphi(a_k) = \varphi(b_k) = y$. Le plus petit des a peut être égal à $-\infty$ et le plus grand des b à $+\infty$. La figure 4.1 donne un exemple d'un tel changement de variable.

	fonction de répartition	densité de probabilité
X	$F(x)$	$f(x)$
$Y=\varphi(X)\ ;\varphi'\geq 0$	$F(x); x=\varphi^{-1}(y)$	$f(x)[\varphi'(x)]^{-1}; x=\varphi^{-1}(y)$
$Y=\varphi(X)\ ;\varphi'<0$	$1-F(x); x=\varphi^{-1}(y)$	$-f(x)[\varphi'(x)]^{-1}; x=\varphi^{-1}(y)$
$Y=X-b$	$F(y+b)$	$f(y+b)$
$Y=aX$	$\begin{cases} F(\frac{y}{a}) & a\geq 0 \\ 1-F(\frac{y}{a}) & a<0 \end{cases}$	$\frac{1}{\lvert a\rvert}f(\frac{y}{a})$

TAB. 4.1: *Fonction de répartition et densité de probabilité de la nouvelle variable aléatoire $Y=\varphi(X)$ où φ est une fonction dérivable.*

On obtient alors la fonction de répartition G de la nouvelle variable aléatoire $Y=\varphi(X)$ par la formule :

$$G(y)=\sum_k F(b_k)-F(a_k);\quad \varphi(a_k)=\varphi(b_k)=y\,. \tag{4.9}$$

La densité de probabilité s'obtient par dérivation, ce qui donne la formule :

$$\boxed{g(y)=\sum_k f(x_k)\frac{1}{\left|\dfrac{d\varphi}{dx}\right|_{x=x_k}}\qquad \varphi(x_k)=y\,.} \tag{4.10}$$

La sommation s'étend sur tous les x_k solutions de l'équation $y=\varphi(x_k)$.

▶**Exemple 4.2.** *Densité de probabilité du carré d'une variable aléatoire.* On a $y=\varphi(x)=x^2$ et $d\varphi/dx=2x$. La variable aléatoire Y étant toujours positive, la densité de probabilité $g(y)$ est nulle en dehors de l'intervalle $y\geq 0$. A un $y\geq 0$ donné, correspondent deux valeurs de x, $x_1=\sqrt{y}$ et $x_2=-\sqrt{y}$, la somme (4.10) s'étend sur ces deux racines x_1 et x_2. Il vient alors :

$$\left|\frac{d\varphi}{dx}\right|_{x=\sqrt{y}}=2\sqrt{y}\quad\text{et}\quad\left|\frac{d\varphi}{dx}\right|_{x=-\sqrt{y}}=2\sqrt{y} \tag{4.11}$$

et l'on obtient ainsi la densité de probabilité de Y :

$$g(y)=\begin{cases}\dfrac{f(\sqrt{y})+f(-\sqrt{y})}{2\sqrt{y}} & y\geq 0\\[2ex] 0 & y<0\end{cases} \tag{4.12}$$

Si la moyenne et la variance existent, et sont respectivement égales à μ et à σ^2, la variable $Y=X^2$ a pour moyenne $\mu^2+\sigma^2$.

▶**Exemple 4.3.** *Densité de probabilité du carré d'une variable aléatoire normale.* Soit X une variable aléatoire normale de densité de probabilité donnée par l'expression :

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}\exp-\frac{(x-\mu)^2}{2\sigma^2} \tag{4.13}$$

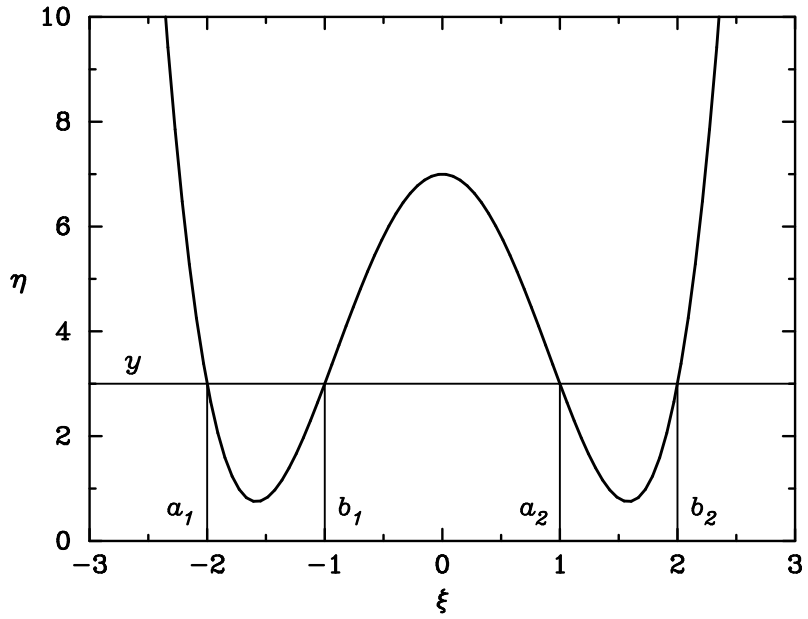


FIG. 4.1: *Exemple de changement de variable aléatoire $Y = \varphi(X)$ continu mais non univoque. Le domaine des X satisfaisant l'inégalité $Y \leq y$ est formé ici des deux segments disjoints $[a_1, b_1] \cup [a_2, b_2]$.*

D'après l'équation (4.12), la densité de probabilité de X^2 est égale à :

$$g(y) = \frac{1}{\sigma\sqrt{2\pi}}\, y^{-\frac{1}{2}} \exp\left(-\frac{y+\mu^2}{2\sigma^2}\right) \cosh\left(\frac{\mu\sqrt{y}}{\sigma^2}\right) . \tag{4.14}$$

Cette loi possède une moyenne $\mu^2 + \sigma^2$ et une variance $4\mu^2\sigma^2 + 2\sigma^4$.

4.1.2 Uniformisation des variables aléatoires continues.

Soit X une variable aléatoire continue de densité de probabilité $f(x)$. Considérons le changement de variable : $Y = F(X)$ où F est la fonction de répartition de X. Par définition de la fonction de répartition, à une valeur x de la variable aléatoire X correspond une valeur y de la variable aléatoire Y donnée par l'expression :

$$y = F(x) = \int_{-\infty}^{x} f(t)\, dt \tag{4.15}$$

Les valeurs possibles de Y sont comprises entre $F(-\infty) = 0$ et $F(\infty) = 1$. La variable aléatoire Y ne pouvant pas prendre de valeurs à l'extérieur de l'intervalle $[0, 1]$, possède donc une densité de probabilité nulle en dehors de cet intervalle. La variable aléatoire X étant continue, F est alors strictement croissante et donc univoque. On a alors pour toutes les valeurs de y comprises entre 0 et 1 :

$$g(y) = f(x)\frac{1}{\dfrac{dF}{dx}} = f(x)\frac{1}{f(x)} = 1 \tag{4.16}$$

Le changement de variable $Y = F(X)$ permet donc de transformer une variable aléatoire continue de densité quelconque, en une variable aléatoire suivant une loi dite uniforme. Par définition la loi uniforme sur $[0, 1]$ possède la densité :

$$f(x) = \begin{cases} 0 & \text{si } x < 0 \text{ ou si } x > 1\,, \\ 1 & \text{si } 0 \leq x \leq 1\,. \end{cases} \tag{4.17}$$

Ce changement de variable est souvent utilisé en traitement d'images, et constitue la méthode dite de *l'égalisation d'histogramme*. On l'utilise également pour générer des nombres aléatoires suivant la loi F quand on a à sa disposition des nombres suivant la loi uniforme. En effet, si les variables aléatoires U_i suivent la loi uniforme, les variables aléatoires $X_i = F^{-1}(U_i)$ suivent la loi F.

▶**Exemple 4.4.** *Simulation de la loi exponentielle.* Une variable aléatoire X suit la loi exponentielle de paramètre $\lambda > 0$ si sa fonction de répartition F est donnée par :

$$F(x) = \begin{cases} 0 & \text{si } x < 0\,, \\ 1 - \exp(-\lambda x) & \text{si } x \geq 0\,. \end{cases} \tag{4.18}$$

D'après ce que nous venons de voir la variable aléatoire $U = F(X)$ suit la loi uniforme sur $[0, 1[$. Réciproquement si l'on dispose d'une variable aléatoire U suivant la loi uniforme sur $[0, 1[$ alors la variable aléatoire $X = F^{-1}(U)$ suivra la loi exponentielle. Il vient : $U = 1 - \exp(-\lambda X)$, $X = -\ln(1 - U)/\lambda$ ou, ce qui revient au même, $X = -\ln(U)/\lambda$ si $U \in]0, 1]$. On dispose ainsi d'un moyen commode pour générer une variable aléatoire exponentielle lorsqu'on dispose d'un générateur de nombres aléatoires suivant la loi uniforme.

4.1.3 Changement de variable et indépendance.

Le théorème suivant limite la classe des changements de variables qui préservent l'indépendance. Dans la pratique il s'agit d'une limitation très peu contraignante.

Théorème 4.1. *(Slutsky.) Si les variables aléatoires $(X_1, \ldots, X_n)$ sont indépendantes et si les fonctions φ_k sont mesurables-Borel, alors les variables aléatoires $\varphi_k(X_k)$ sont aussi indépendantes.*

On dit qu'une fonction φ est mesurable-Borel si l'ensemble des x défini par $g(x) < c$ pour tout $c \in \mathbb{R}$, est un borélien. Une fonction continue est mesurable-Borel.

On trouvera la démonstration de ce théorème au chapitre 4 §5 de l'ouvrage de Rényi [62].

▶**Exemple 4.5.** Si les variables aléatoires indépendantes $(X_1, \ldots, X_n)$ sont transformées par élévation à la puissance r, alors les variables aléatoires $(X_1^r, \ldots, X_n^r)$ sont aussi indépendantes car le changement de variable $x \mapsto x^r$ est continu.

4.2 Plusieurs fonctions de plusieurs variables.

Au vecteur aléatoire $\boldsymbol{X}$ de composantes $(X_1, \ldots, X_n)$, on fait correspondre un autre vecteur aléatoire $\boldsymbol{Y}$ de composantes $(Y_1, \ldots, Y_n)$, par l'intermédiaire

de n fonctions φ_i, telles que :

$$\begin{aligned} Y_1 &= \varphi_1(X_1, \ldots, X_n)\,, \\ Y_2 &= \varphi_2(X_1, \ldots, X_n)\,, \\ &\ldots\ldots\ldots\ldots\ldots\ldots \\ Y_n &= \varphi_n(X_1, \ldots, X_n)\,. \end{aligned} \tag{4.19}$$

Pour un $y : (y_1, \ldots, y_n)$ donné, on notera $x^{(k)} : (x_1^{(k)}, \ldots, x_n^{(k)})$ les solutions du système $y_i = \varphi_i(x_1^{(k)}, \ldots, x_n^{(k)})\ \forall i,\ 1 \leq i \leq n$. Par un raisonnement analogue à celui fait sur le changement d'une seule variable aléatoire, on trouverait la formule générale donnant la densité de probabilité de $\boldsymbol{Y}$:

$$\boxed{g(y_1, \ldots, y_n) = \sum_k f(x_1^{(k)}, \ldots, x_n^{(k)}) \left| \frac{\partial(y_1, \ldots, y_n)}{\partial(x_1, \ldots, x_n)} \right|^{-1}_{x=x^{(k)}}} \tag{4.20}$$

L'expression $\partial(y_1, \ldots, y_n)/\partial(x_1, \ldots, x_n)$, est le jacobien J du changement de variables.

▶**Exemple 4.6.** *Génération de nombres pseudo-aléatoires suivant la loi normale réduite.* Soient deux variables aléatoires indépendantes U_1 et U_2 suivant la loi uniforme sur $]0,1]\times]0,1]$. A partir du couple (U_1, U_2), définissons le nouveau couple (X_1, X_2) par l'intermédiaire du changement de variables :

$$\begin{aligned} x_1 &= \sqrt{-2\ln u_1}\cos(2\pi u_2)\,, \\ x_2 &= \sqrt{-2\ln u_1}\sin(2\pi u_2)\,. \end{aligned} \tag{4.21}$$

Ce changement de variables est bijectif. En effet on obtient la fonction inverse en posant : $r = \sqrt{x_1^2 + x_2^2}$ et $2\pi u = \arccos(x_1/r)$, il vient alors :

$$\begin{aligned} u_1 &= \exp -\tfrac{1}{2}r^2\,, \\ u_2 &= \begin{cases} u & \text{si } x_2 \geq 0, \\ 1-u & \text{si } x_2 < 0. \end{cases} \end{aligned} \tag{4.22}$$

On a $r = \sqrt{-2\ln u_1}$, et le jacobien du changement de variables (4.21) est :

$$J = \begin{vmatrix} \dfrac{\partial x_1}{\partial u_1} & \dfrac{\partial x_1}{\partial u_2} \\ \dfrac{\partial x_2}{\partial u_1} & \dfrac{\partial x_2}{\partial u_2} \end{vmatrix} = \begin{vmatrix} \dfrac{1}{u_1 r}\cos(2\pi u_2) & -2\pi r\sin(2\pi u_2) \\ \dfrac{1}{u_1 r}\sin(2\pi u_2) & 2\pi r\cos(2\pi u_2) \end{vmatrix} = \frac{2\pi}{u_1}\,.$$

Il vient :

$$g(x_1, x_2) = f(u_1, u_2)|J|^{-1} = \frac{u_1}{2\pi} = \frac{1}{2\pi}\exp\{-\tfrac{1}{2}(x_1^2 + x_2^2)\}\,. \tag{4.23}$$

Nous avons donc démontré que le changement de variables (4.21) transforme un couple de variables aléatoires uniformes indépendantes, en un couple de variables aléatoires normales réduites. Les variables aléatoires X_1 et X_2 sont elles-mêmes indépendantes car leur loi conjointe $g(x_1, x_2)$ peut être mise sous la forme du produit de ses lois marginales :

$$g(x_1, x_2) = \frac{1}{\sqrt{2\pi}}\exp(-\tfrac{1}{2}x_1^2)\frac{1}{\sqrt{2\pi}}\exp(-\tfrac{1}{2}x_2^2)\,.$$

4.3 Une fonction de plusieurs variables.

Soient n variables aléatoires $(X_1, \ldots, X_n)$ et une seule fonction φ définissant une nouvelle variable aléatoire $Y = \varphi(X_1, \ldots, X_n)$. Pour connaître la densité de probabilité de Y, on est amené à introduire les variables aléatoires Y_i telles que :

$$\begin{aligned} Y_1 &= X_1 \,, \\ Y_2 &= X_2 \,, \\ &\cdots\cdots\cdots \\ Y_{n-1} &= X_{n-1} \,, \\ Y_n &= Y = \varphi(X_1, \ldots, X_n) \,. \end{aligned} \tag{4.24}$$

La valeur absolue du jacobien de ce changement de variables est égale à :

$$|J| = \left| \frac{\partial \varphi}{\partial x_n} \right| , \tag{4.25}$$

elle nous permet d'obtenir g_n la densité de probabilité conjointe des variables aléatoires $Y_1, \ldots, Y_n$:

$$g_n(y_1, \ldots, y_{n-1}, y) = \sum_k f(x_1^{(k)}, \ldots, x_n^{(k)}) \left| \frac{\partial \varphi}{\partial x_n} \right|^{-1}_{x = x^{(k)}} , \tag{4.26}$$

à partir de laquelle on obtient la densité de probabilité de Y comme densité de probabilité marginale :

$$g(y) = \int_{-\infty}^{\infty} g_n(y_1, \ldots, y_{n-1}, y)\, dy_1 \cdots dy_{n-1} \,. \tag{4.27}$$

Nous allons appliquer ces formules dans trois cas importants.

4.3.1 Somme et différence de deux variables aléatoires.

Pour la somme on a $Y = X_1 + X_2$. On suppose connue la densité de probabilité conjointe $f_2(x_1, x_2)$ du couple aléatoire (X_1, X_2). On pose donc $Y_1 = X_1$ et $Y_2 = X_1 + X_2$. Cette transformation a pour jacobien :

$$J = \frac{\partial(y_1, y_2)}{\partial(x_1, x_2)} = \begin{vmatrix} 1 & 0 \\ 1 & 1 \end{vmatrix} = 1 \,. \tag{4.28}$$

En remplaçant la valeur du jacobien dans l'équation (4.20), on trouve :

$$g_2(y_1, y_2) = f_2(y_1, y_2 - y_1) \,.$$

La densité de probabilité de Y est égale à la densité marginale :

$$g(y) = \int_{-\infty}^{\infty} f_2(u, y - u)\, du \,. \tag{4.29}$$

Si les variables aléatoires X_1, X_2 sont indépendantes, la densité du couple est égale au produit de ses densités marginales, on a $f_2(x_1, x_2) = f_{X_1}(x_1) f_{X_2}(x_2)$ et la formule précédente s'écrit :

$$g(y) = \int_{-\infty}^{\infty} f_{X_1}(u) f_{X_2}(y - u)\, du \,. \tag{4.30}$$

C'est le produit de convolution de f_{X_1} par f_{X_2}. Pour la différence $Y = X_2 - X_1$, on obtiendrait de la même façon :

$$g(y) = \int_{-\infty}^{\infty} f_{X_1}(u) f_{X_2}(y+u)\, du\,. \tag{4.31}$$

4.3.2 Produit de deux variables aléatoires.

On a $Y = X_1 X_2$, et on suppose connue la densité de probabilité conjointe $f_2(x_1, x_2)$ du couple aléatoire (X_1, X_2). On pose alors $Y_1 = X_1$ et $Y_2 = X_1 X_2$. Cette transformation a pour jacobien :

$$J = \frac{\partial(y_1, y_2)}{\partial(x_1, x_2)} = \begin{vmatrix} 1 & 0 \\ x_2 & x_1 \end{vmatrix} = |x_1|\,. \tag{4.32}$$

En remplaçant la valeur du jacobien dans l'équation (4.20), on trouve :

$$g_2(y_1, y_2) = f_2(y_1, y_2/y_1)\frac{1}{|y_1|}\,.$$

La densité de probabilité de Y est la densité marginale :

$$g(y) = \int_{-\infty}^{\infty} f_2\left(u, \frac{y}{u}\right)\frac{1}{|u|}\, du\,. \tag{4.33}$$

Si les variables aléatoires X_1, X_2 sont indépendantes, on a :

$$g(y) = \int_{-\infty}^{\infty} f_{X_1}(u) f_{X_2}\left(\frac{y}{u}\right)\frac{1}{|u|}\, du\,. \tag{4.34}$$

▶**Exemple 4.7.** *Densité de probabilité du produit de deux variables aléatoires normales indépendantes.* Soient X_1, X_2 deux variables aléatoires normales de moyennes nulles, de variances $\sigma_1 = \sigma_2 = 1$. On trouve la densité de probabilité $g(y)$ de leur produit $Y = X_1 X_2$ en appliquant la formule (4.34). Il vient :

$$\begin{aligned} g(y) &= \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} \exp\{-\tfrac{1}{2}u^2\} \frac{1}{\sqrt{2\pi}} \exp\{-\tfrac{1}{2}\frac{y^2}{u^2}\}\, \frac{du}{|u|}\,, \\ &= \frac{1}{\pi}\int_0^{+\infty} \exp\{-\tfrac{1}{2}[u^2 + \frac{y^2}{u^2}]\}\, \frac{du}{u}\,. \end{aligned}$$

En posant $e^t = \dfrac{u^2}{|y|}$ il vient :

$$\begin{aligned} g(y) &= \frac{1}{\pi}\int_{-\infty}^{+\infty} \exp\{-\tfrac{1}{2}|y|(e^t + e^{-t})\}\, \frac{dt}{2} = \frac{1}{\pi}\int_0^{+\infty} \exp\{-|y|\cosh t\}\, dt\,, \\ &= \frac{1}{\pi} K_0(|y|)\,, \end{aligned}$$

où K_0 est la fonction de Bessel modifiée de 2$^{\text{e}}$ espèce et d'ordre 0. Cette densité possède une moyenne nulle et un écart type égal à 1, son graphe est présenté sur la figure 4.2.

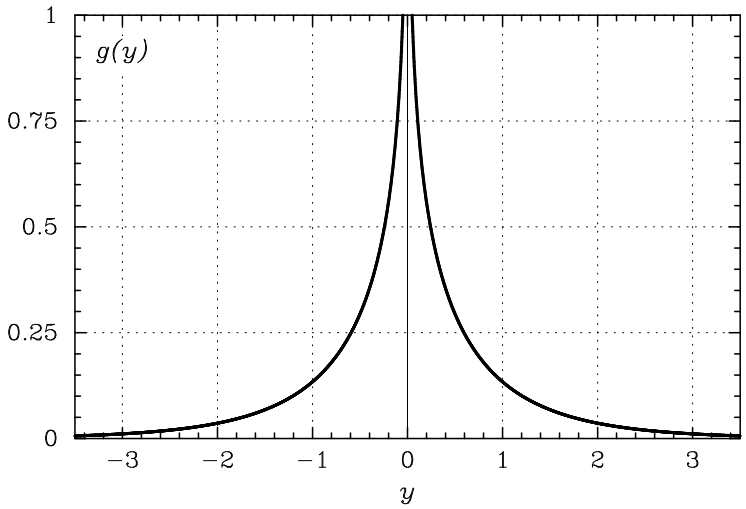


FIG. 4.2: *Densité de probabilité $g(y)$ du produit de deux variables aléatoires normales réduites X_1 et X_2. On a $\mu_1 = \mu_2 = 0$, $\sigma_1 = \sigma_2 = 1$, $Y = X_1 X_2$ et $g(y) = \frac{1}{\pi} K_0(|y|)$ où K_0 est la fonction de Bessel modifiée de deuxième espèce et d'ordre 0.*

4.3.3 Quotient de deux variables aléatoires.

On a $Y = X_2/X_1$ et on suppose connue la densité de probabilité conjointe $f_2(x_1, x_2)$ du couple aléatoire (X_1, X_2). On pose alors $Y_1 = X_1$ et $Y_2 = X_2/X_1$. Cette transformation a pour jacobien :

$$J = \frac{\partial(y_1, y_2)}{\partial(x_1, x_2)} = \begin{vmatrix} 1 & 0 \\ -\dfrac{x_2}{x_1^2} & \dfrac{1}{x_1} \end{vmatrix} = \left| \frac{1}{x_1} \right| . \tag{4.35}$$

En remplaçant la valeur du jacobien dans l'équation (4.20), on trouve :

$$g_2(y_1, y_2) = f_2(y_1, y_1 y_2)|y_1| .$$

La densité de probabilité de Y est la densité marginale :

$$g(y) = \int_{-\infty}^{\infty} f_2(u, uy)|u| \, du . \tag{4.36}$$

Si les variables aléatoires X_1, X_2 sont indépendantes on a :

$$g(y) = \int_{-\infty}^{\infty} f_{X_1}(u) f_{X_2}(uy)|u| \, du . \tag{4.37}$$

Ces propriétés sont résumées dans le tableau 4.2.

▶**Exemple 4.8.** *Densité de probabilité du quotient de deux variables aléatoires normales.* Soient X_1, X_2 deux variables aléatoires normales réduites indépendantes. En tant que variables réduites on a : $\mu_1 = \mu_2 = 0$ et $\sigma_1 = \sigma_2 = 1$. En application de la

	Variables quelconques.	*Variables indépendantes.*
$X_2 \pm X_1$	$g(y) = \int_{-\infty}^{\infty} f(u, y \mp u)\, du$	$g(y) = \int_{-\infty}^{\infty} f_{X_1}(u) f_{X_2}(y \mp u)\, du$
$X_1 \times X_2$	$g(y) = \int_{-\infty}^{\infty} f\left(u, \frac{y}{u}\right) \frac{1}{\lvert u\rvert}\, du$	$g(y) = \int_{-\infty}^{\infty} f_{X_1}(u) f_{X_2}\left(\frac{y}{u}\right) \frac{1}{\lvert u\rvert}\, du$
$\frac{X_2}{X_1}$	$g(y) = \int_{-\infty}^{\infty} f(u, uy)\lvert u\rvert\, du$	$g(y) = \int_{-\infty}^{\infty} f_{X_1}(u) f_{X_2}(uy)\lvert u\rvert\, du$

TAB. 4.2: *Densité de probabilité $g(y)$ d'une variable aléatoire η égale à l'une des quatre opérations sur le couple de variables aléatoires (X_1, X_2). La fonction f désigne la densité de probabilité conjointe du couple (X_1, X_2), les fonctions f_{X_1} et f_{X_2}, désignent les densités de X_1 et X_2.*

formule (4.37), on trouve la densité de probabilité du quotient X_1/X_2 :

$$\begin{aligned} g(y) &= \frac{1}{2\pi} \int_{-\infty}^{+\infty} e^{-\frac{1}{2}u^2} e^{-\frac{1}{2}(uy)^2} |u|\, du\,, \\ &= \frac{1}{2\pi} \int_{-\infty}^{+\infty} e^{-\frac{1}{2}u^2(1+y^2)} |u|\, du\,, \\ &= \frac{1}{\pi} \int_{0}^{+\infty} e^{-t(1+y^2)}\, dt\,, \\ &= \frac{1}{\pi} \frac{1}{1+y^2}\,. \end{aligned}$$

La loi suivie par le rapport de deux variables aléatoires normales réduites est une loi de *Cauchy*, c'est-à-dire une loi qui ne possède pas de moments (et en particulier pas de moyenne).

4.4 Le point de vue des probabilités conditionnelles.

On aurait également pu développer la théorie du changement de variable aléatoire dans le cadre plus formel des probabilités conditionnelles. Illustrons cette démarche dans le cas d'une fonction d'une variable aléatoire. Soit $Y = \varphi(X)$ le changement de variable et $f(x)$ la densité de X. La fonction de répartition $G(y)$ de Y s'obtient en sommant la densité f sur le domaine des x où $\varphi(x) \leq y$; il s'agit donc bien d'une fonction de répartition conditionnelle. Soit $\Omega'(y)$ ce domaine. On a alors :

$$G(y) = \int_{\Omega'(y)} f(x)\, dx\,. \tag{4.38}$$

Introduisons la distribution de Heaviside $H(y - \varphi(x))$ qui vaut 0 si $y < \varphi(x)$ et 1 si $y \geq \varphi(x)$. A l'aide de cette distribution, l'équation précédente peut se mettre sous la forme :

$$G(y) = \int_{\Omega} f(x) H(y - \varphi(x))\, dx\,, \tag{4.39}$$

où la sommation s'étend maintenant à tout l'espace Ω de définition de X. La densité de probabilité y s'obtient par dérivation de $G(y)$, et en notant que la dérivée de la distribution de Heaviside H est égale à la distribution de Dirac δ de même argument on a :

$$g(y) = \int_\Omega f(x)\delta(y - \varphi(x))\, dx\,. \tag{4.40}$$

En sachant que $\delta(h(x))$ se calcule à l'aide de la formule :

$$\delta(h(x)) = \sum_i \frac{1}{|h'(x)|}_{x=x_i} \delta(x - x_i)\,, \tag{4.41}$$

où la sommation s'étend sur l'ensemble des x_i tels que $h(x_i) = 0$. Ici $h(x) = y - \varphi(x)$, et l'on obtient finalement :

$$g(y) = \sum_i \frac{f(x_i)}{|\varphi'(x)|}_{x=x_i}\,. \tag{4.42}$$

La sommation s'étend cette fois sur l'ensemble des x_i tels que $\varphi(x_i) = y$ et qui appartiennent au domaine de définition Ω de la variable aléatoire X. On retrouve ainsi la formule (4.7). La démarche est identique, pour le cas de plusieurs variables aléatoires.

4.5 Exemples.

4.5.1 Module et phase d'un couple de variables aléatoires

On considère deux variables aléatoires X_1 et X_2, on note x_1 et x_2 les valeurs prises par ces variables aléatoires et $f(x_1, x_2)$ la densité de probabilité de ce couple. Le module R et la phase Φ sont les coordonnées polaires associées au couple de coordonnées cartésiennes X_1 et X_2. Le changement de variables est donc :

$$\begin{aligned} x_1 &= r\cos\phi, \\ x_2 &= r\sin\phi. \end{aligned} \tag{4.43}$$

Ce changement de variable est bijectif (voir exemple 4.6) sauf en $r = 0$ qui est un ensemble de mesure nulle. Le jacobien du changement de variables de (r, ϕ) vers (x_1, x_2) est égal à r. On trouve alors la densité de probabilité $g(r, \phi)$ du couple de variables aléatoires (R, Φ) connaissant la densité de probabilité $f(x_1, x_2)$ du couple (X_1, X_2) grâce à la formule (4.20). Il vient :

$$g(r, \phi) = rf(x_1, x_2) = rf(r\cos\phi, r\sin\phi)\,. \tag{4.44}$$

Les densités de la phase et du module sont les densités marginales de la loi du couple (R, Φ). Nous allons appliquer ce calcul au cas de deux variables aléatoires normales indépendantes.

4.5.2 Module et phase d'un couple de variables aléatoires normales indépendantes.

Les deux variables aléatoires indépendantes X_1 et X_2 suivent respectivement une loi normale de paramètres μ_1, σ_1 et une loi normale de paramètres μ_2, σ_2. La loi du couple (R, Φ) s'obtient par (4.44), il vient :

$$\begin{aligned} g(r,\phi) &= r\frac{1}{\sqrt{2\pi}\sigma_1}\exp\left\{-\frac{(r\cos\phi-\mu_1)^2}{2\sigma_1^2}\right\}\frac{1}{\sqrt{2\pi}\sigma_2}\exp\left\{-\frac{(r\sin\phi-\mu_2)^2}{2\sigma_2^2}\right\}, \\ &= \frac{r}{2\pi\sigma_1\sigma_2}\exp-\left\{\frac{(r\cos\phi-\mu_1)^2}{2\sigma_1^2}+\frac{(r\sin\phi-\mu_2)^2}{2\sigma_2^2}\right\}. \end{aligned} \tag{4.45}$$

Densité de probabilité du module (loi de Rayleigh-Rice).

Cette densité s'obtient en intégrant la phase ϕ sur tout son domaine de définition. Supposons pour simplifier que $\sigma_1 = \sigma_2 = \sigma$, il vient :

$$\begin{aligned} g_R(r) &= \frac{r}{2\pi\sigma^2}\int_0^{2\pi}\exp\left\{-\frac{(r\cos\phi-\mu_1)^2+(r\sin\phi-\mu_2)^2}{2\sigma^2}\right\}d\phi\,, \\ &= \frac{r}{2\pi\sigma^2}\exp\left(-\frac{r^2+\mu^2}{2\sigma^2}\right)\int_0^{2\pi}\exp\left\{-\frac{r\mu}{\sigma^2}\cos(\phi-\phi_0)\right\}d\phi\,, \end{aligned}$$

où μ et ϕ_0 sont les coordonnées polaires du couple (μ_1, μ_2), c'est-à-dire : $\mu = \sqrt{\mu_1^2+\mu_2^2}$ et $\mu_1 = \mu\cos\phi_0$, $\mu_2 = \mu\sin\phi_0$. La fonction à intégrer étant périodique il vient :

$$\frac{1}{2\pi}\int_0^{2\pi}\exp\left\{-\frac{r\mu}{\sigma^2}\cos(\phi-\phi_0)\right\}d\phi = \frac{1}{2\pi}\int_0^{2\pi}\exp\left(-\frac{r\mu}{\sigma^2}\cos\phi\right)d\phi = I_0\left(\frac{r\mu}{\sigma^2}\right).$$

Dans cette expression I_0 représente la fonction de Bessel modifiée d'ordre 0 (voir par exemple la formule 8.431.3 de Gradshteyn et Ryzhik [27]). La densité de probabilité du module R de deux variables aléatoires normales indépendantes et de même écart type σ est alors donnée par l'expression :

$$g_R(r) = \frac{r}{\sigma^2}\exp\left(-\frac{r^2+\mu^2}{2\sigma^2}\right)I_0\left(\frac{r\mu}{\sigma^2}\right), \quad r \geq 0, \tag{4.46}$$

où μ est la moyenne quadratique des moyennes μ_1 et μ_2 des deux variables. Cette densité de probabilité porte le nom de densité de Rayleigh-Rice. Si μ est grand devant σ, la moyenne μ_R et la variance σ_R^2 de la loi de Rayleigh-Rice sont données par les formules asymptotiques suivantes :

$$\mu_R \approx \mu\left(1+\frac{\sigma^2}{2\mu^2}\right), \qquad \sigma_R^2 \approx \sigma^2\left(1-\frac{\sigma^2}{4\mu^2}\right). \tag{4.47}$$

L'expression exacte des moments non-centrés μ'_k de R est donnée au chapitre 3.2.2 de l'ouvrage de Lévine [48], elle vaut :

$$\mu'_k = (2\sigma^2)^{\frac{k}{2}}\Gamma(1+\tfrac{k}{2})\,{}_1F_1(-\tfrac{k}{2}, 1, -\tfrac{\mu^2}{2\sigma^2}), \tag{4.48}$$

où Γ et ${}_1F_1$ représentent respectivement la fonction eulerienne de seconde espèce et une fonction hypergéométrique dégénérée (voir appendice A.1).

Densité de probabilité de la phase.

Nous nous plaçons toujours dans le cas où les variables aléatoires X_1 et X_2 possèdent le même écart type σ. La loi suivie par la phase s'obtient en intégrant l'expression (4.45) sur le domaine $r \geq 0$ avec $\sigma_1 = \sigma_2 = \sigma$. Il vient :

$$\begin{aligned} g_\Phi(\phi) &= \frac{1}{2\pi\sigma^2} \int_0^\infty r \exp\left\{ -\frac{(r\cos\phi - \mu_1)^2 + (r\sin\phi - \mu_2)^2}{2\sigma^2} \right\} dr, \\ &= \frac{1}{2\pi\sigma^2} \exp\left(-\frac{\mu^2}{2\sigma^2} \right) \int_0^\infty r \exp\left\{ -\frac{r^2 - 2r\mu\cos(\phi - \phi_0)}{2\sigma^2} \right\} dr \,. \end{aligned}$$

En complètant la forme quadratique apparaissant dans l'argument de la seconde exponentielle de façon à former un carré parfait on obtient :

$$g_\Phi(\phi) = \frac{1}{2\pi\sigma^2} \exp\{-\frac{\mu^2}{2\sigma^2}\sin^2(\phi - \phi_0)\} \int_0^\infty r \exp\left\{ -\frac{(r - \mu\cos(\phi - \phi_0))^2}{2\sigma^2} \right\} dr \,.$$

En posant $r_0 = \mu\cos(\phi - \phi_0)$, il vient :

$$\begin{aligned} g_\Phi(\phi) &= \frac{1}{2\pi\sigma^2} \exp\{-\frac{\mu^2 - r_0^2}{2\sigma^2}\} \int_0^\infty r \exp\left\{ -\frac{(r - r_0)^2}{2\sigma^2} \right\} dr \,, \\ &= \frac{1}{2\pi\sigma^2} \exp\{-\frac{\mu^2 - r_0^2}{2\sigma^2}\} \int_{-r_0}^\infty (r + r_0) \exp -\frac{r^2}{2\sigma^2} \, dr \,. \end{aligned}$$

L'intégrale se scinde en deux parties :

$$\begin{aligned} \int_{-r_0}^\infty (r + r_0) \exp\{-\frac{r^2}{2\sigma^2}\} \, dr &= \int_{|r_0|}^\infty r \exp\{-\frac{r^2}{2\sigma^2}\} \, dr + r_0 \int_{-\infty}^{r_0} \exp\{-\frac{r^2}{2\sigma^2}\} \, dr \\ &= \sigma^2 \exp\{-\frac{r_0^2}{2\sigma^2}\} + \sqrt{2\pi}\sigma r_0 \Phi(\frac{r_0}{\sigma}) \,, \end{aligned}$$

où Φ est la fonction de répartition de la loi normale réduite (Fonction de Laplace). Finalement :

$$g_\Phi(\phi) = \frac{1}{2\pi} \exp\{-\frac{\mu^2}{2\sigma^2}\} + \frac{r_0}{\sqrt{2\pi}\sigma} \Phi(\frac{r_0}{\sigma}) \exp\{-\frac{\mu^2 - r_0^2}{2\sigma^2}\} \,,$$

soit en revenant à la variable ϕ :

$$\begin{aligned} g_\Phi(\phi) = \frac{1}{2\pi} \exp\{-\frac{\mu^2}{2\sigma^2}\} + \\ \frac{\mu\cos(\phi - \phi_0)}{\sqrt{2\pi}\sigma} \Phi\{\frac{\mu\cos(\phi - \phi_0)}{\sigma}\} \exp\{-\frac{\mu^2\sin^2(\phi - \phi_0)}{2\sigma^2}\} \,. \end{aligned} \quad (4.49)$$

La fonction g_Φ est périodique de période 2π, cependant on ne doit considérer qu'une seule période, par exemple : $\phi \in [0, 2\pi[$ ou $\phi - \phi_0 \in [-\pi, \pi[$. Sur la période $\phi - \phi_0 \in [-\pi, \pi[$, la fonction g_Φ est symétrique et tous les moments pairs de la variable aléatoire $\Phi - \phi_0$ sont nuls. Si $\mu = 0$ la fonction g_Φ est constante et la variable aléatoire associée suit alors une loi uniforme sur la période considérée. Dans ce cas les variables aléatoires R et Φ sont indépendantes. Si μ est grand devant σ la variable $\Phi - \phi_0$ suit approximativement une loi normale de moyenne nulle et de variance $(\sigma/\mu)^2$. On trouvera une étude plus détaillée de cette loi au Chapitre 3 §2.3 de l'ouvrage de Lévine [48].

4.6 Aspects numériques.

Le programme **RUNIF** suivant génère une série de nombre pseudo-aléatoires suivant la loi uniforme entre $[0, 1[$. L'utilisateur doit fournir au départ trois entiers : **IX, IY** et **IZ** qui doivent être des nombres premiers assez grands, si **IX** est négatif ou nul le programme fournira ces trois nombres. La partie intitulée « calcul » est inspirée de Wichmann et Hill (1982) [71]. On consultera aussi James (1990) [35] pour avoir un aperçu sur d'autres méthodes classiques permettant de générer des nombres pseudo-aléatoires suivant la loi uniforme.

```
      REAL*4 FUNCTION RUNIF(IX,IY,IZ)
      INTEGER*4 IX, IY, IZ
      INTEGER*4 SEED(3)/15251,25159,14981/
!   Initialisation
      IF ( IX.LE.0 ) THEN
       IX = ISEED(1)
       IY = ISEED(2)
       IZ = ISEED(3)
      ENDIF
!   Calcul
      IX = MOD(IX*171,30269)
      IY = MOD(IY*172,30307)
      IZ = MOD(IZ*170,30323)
      RUNIF = (FLOAT(IX)/30269 + FLOAT(IY)/30307 + FLOAT(IZ)/30323)
      RUNIF = RUNIF - JINT(RUNIF)
      RETURN
      END
```

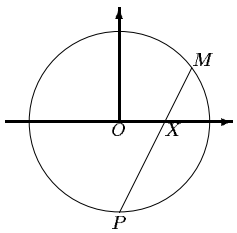
Ce programme (ou un programme équivalent) servira de base pour simuler d'autres lois à l'aide de la méthode du changement de variable. Par exemple, si $X = -\ln(\texttt{runif(ix,iy,iz)})$ la variable X suivra la loi exponentielle de paramètre $\lambda = 1$.

4.7 Exercices et problèmes.

On trouvera au chapitre 8 la définition des lois non encore introduites à ce niveau.

Exercice 4.1. Une variable aléatoire Θ suit une loi de densité de probabilité uniforme entre 0 et 2π : $f(\theta) = \frac{1}{2\pi}$. Trouver l'expression de la densité de probabilité des variables aléatoires $X = \cos\Theta$, $Y = \sin\Theta$ et de $Z = \tan\Theta$.

Exercice 4.2. *Projection stéréographique.* Soit un cercle de centre O et un point fixe P du cercle. La projection stéréographique de P est le point X intersection de la droite PM avec l'axe perpendiculaire au rayon OP (voir figure). Montrer que si M est réparti uniformément sur le cercle alors la variable aléatoire $\xi = OX$ suit une loi de Cauchy.



Exercice 4.3. Soient 2 variables aléatoires indépendantes X_1 et X_2 suivant chacune la loi normale de moyenne nulle et de variance σ^2. Trouver la fonction de répartition et la densité de probabilité des variables : $|X_1|$, $|X_1| + |X_2|$, $|X_1| + X_2$ et $|X_1| - |X_2|$.

Exercice 4.4. *Densité du produit de deux variables aléatoires normales dépendantes.* Soient X_1, X_2 deux variables aléatoires normales de moyennes nulles, de variances σ_1, σ_2 et de coefficient de corrélation ρ (voir la définition (6.11), page 84).

Montrer que la densité de probabilité $g(y)$ du produit $Y = X_1X_2$ de ces variables est donnée par l'expression :

$$g(y) = \frac{1}{\pi\sigma_1\sigma_2(1-\rho^2)^{\frac{1}{2}}} K_0\Big(\frac{|y|}{\sigma_1\sigma_2(1-\rho^2)}\Big) \exp\Big(\frac{\rho y}{\sigma_1\sigma_2(1-\rho^2)}\Big), \tag{4.50}$$

où K_0 est la fonction de Bessel modifiée de 2$^{\text{e}}$ espèce et d'ordre 0.

Exercice 4.5. Trouver la densité de probabilité du produit de 2 variables aléatoires indépendantes et uniformément réparties sur le segment $[a, b]$ tel que $0 < a < b$.

Exercice 4.6. Soient n variables aléatoires indépendantes $U_1, \ldots, U_n$ suivant la loi uniforme entre -1 et 1. Montrer que la densité de probabilité $f_3(x)$ de la loi suivie par la somme de trois de ces variables est égale à :

$$f_3(x) = \begin{cases} 0 & \text{si } |x| \geq 3\,, \\ \dfrac{(3-|x|)^2}{16} & \text{si } 1 \leq |x| \leq 3\,, \\ \dfrac{3-x^2}{8} & \text{si } 0 \leq |x| \leq 1\,. \end{cases} \tag{4.51}$$

Montrer par récurrence que la densité de probabilité $f_n(x)$ de la loi suivie par la somme de n variables aléatoires U_i mutuellement indépendantes, est donnée par la formule :

$$f_n(x) = \begin{cases} \dfrac{1}{2^n(n-1)!} \displaystyle\sum_{k=0}^{\lfloor \frac{1}{2}(n+x)\rfloor} (-1)^k C_n^k (n+x-2k)^{n-1} & \text{si } |x| < n\,, \\ 0 & \text{si } |x| \geq n\,, \end{cases} \tag{4.52}$$

où $\lfloor x \rfloor$ désigne la partie entière de x (c'est-à-dire le plus grand entier inférieur ou égal à x).

Exercice 4.7. Soient $U_1, \ldots, U_n$ des variables aléatoires indépendantes suivant la loi uniforme entre 0 et 1. Montrer que la variable aléatoire $X = -2\ln(U_1U_2\cdots U_n)$ suit une loi du χ^2 a $2n$ degrés de liberté.

Exercice 4.8. On désire répartir des points uniformément sur une sphère. Par uniforme, on entend que la probabilité d'obtenir un point dans l'angle solide $d\Omega$ est égale à $\frac{d\Omega}{4\pi}$. On dispose d'un générateur de nombres aléatoires fournissant une suite ($U_1, \ldots,$-U_n) de variables aléatoires indépendantes suivant la loi uniforme entre 0 et 1. Trouver un changement de variable sur les membres U_i de la suite de façon à obtenir le résultat souhaité.

Problème 4.9. Soient 4 variables aléatoires X_1, X_2, X_3, X_4 indépendantes et suivant toutes la loi exponentielle de paramètre $\lambda = 1$ (voir exemple 4.4). On considère le changement de variables :

$$\begin{aligned} Q_1 &= X_1 + X_2 + X_3 + X_4 \\ Q_1Q_2 &= X_2 + X_3 + X_4 \\ Q_1Q_2Q_3 &= X_3 + X_4 \\ Q_1Q_2Q_3Q_4 &= X_4\,. \end{aligned} \tag{4.53}$$

Calculer la densité de probabilité conjointe suivie par les 4 variables Q_1, Q_2, Q_3, Q_4. Montrer que les variables aléatoires Q_1, Q_2, Q_3 et Q_4 sont indépendantes et donner leur densité de probabilité.

En s'inspirant de l'exemple 4.4 donner les changements de variables permettant de générer des nombres pseudo-aléatoires suivant les lois de Q_2, Q_3 et Q_4. Trouver egalement comment simuler la loi de Q_1.

Montrer que la densité conditionnelle $f(q_2, q_3, q_4|Q_1 = q_1)$ du triplet (Q_2, Q_3, Q_4) sachant que $Q_1 = q_1 > 0$ est uniforme sur un domaine de définition que l'on précisera.

Problème 4.10. *Distribution des nombres.* En base b, un nombre réel quelconque X peut toujours être mis sous la forme d'un produit d'un nombre compris entre $1/b$ et 1, par une puissance entière de b. le premier nombre est appelé la « mantisse » et le deuxième «l'ordre de grandeur». Par exemple en base 10, $b = 10$ et le nombre $X = 0.00123$ a pour mantisse : 0.123 et pour ordre de grandeur : 10^{-2}.

Montrer que si la mantisse de X suit la loi réciproque dite « loi de Benford » de densité de probabilité :

$$r(x) = \frac{1}{\ln b}\frac{1}{x}, \quad \frac{1}{b} \leq x < 1\,,$$

alors la mantisse de $Z = XY$ suit également la loi réciproque, et cela quelle que soit la loi suivie par Y. Montrer que cette propriété est aussi vraie pour les divisions $Z = X/Y$ et $Z = Y/X$ (Hamming, 1970, [28]).

Chapitre 5

Nombres et fonctions caractéristiques.

5.1 L'espérance mathématique.

5.1.1 L'espérance mathématique des variables aléatoires discrètes.

L'espérance mathématique d'une variable aléatoire discrète X, c'est-à-dire d'une variable aléatoire dont les valeurs possibles x_i sont dénombrables, est la moyenne arithmétique de ces valeurs pondérées par leur probabilité. Soient x_i, $i \in \mathbb{N}$ les valeurs possibles de X et $p_i = \Pr\{X = x_i\}$ la probabilité avec laquelle elles apparaissent. L'*espérance mathématique* $\mathrm{E}\{X\}$ de X est définie par :

$$\mathrm{E}\{X\} = \sum_{i \in \mathbb{N}} x_i \, p_i \, . \tag{5.1}$$

Cette somme peut être finie ou infinie. Dans le cas où elle est infinie elle peut diverger comme le montre le cas où $\Pr\left\{X = 2^i\right\} = 2^{-i}$, $i \in \mathbb{N}$ et pour lequel $\sum_i x_i p_i = \infty$. Dans le cas où la série (5.1) converge, on exige de plus qu'elle converge quel que soit l'ordre dans lequel on effectue la sommation. L'ordre de sommation de la série dépend de la façon dont on « numérote » les issues de X à l'aide de l'indice i et il est alors légitime de vouloir trouver le même résultat indépendamment de tel ou tel numérotage particulier. Une condition nécessaire et suffisante pour qu'une série converge quel que soit l'ordre de ses termes est qu'elle converge absolument. En conséquence, l'espérance mathématique $\mathrm{E}\{X\}$ n'est définie par (5.1) que si la série pondérée des $|x_i|$ converge, soit :

$$\mathrm{E}\{X\} = \sum_{i \in \mathbb{N}} x_i \, p_i, \quad \text{si} \quad \sum_{i \in \mathbb{N}} |x_i| \, p_i < \infty \, . \tag{5.2}$$

Dans la suite de cet exposé, il sera sous-entendu que toute définition impliquant une série semblable à (5.1) sera subordonnée à la convergence absolue de cette série.

▶**Exemple 5.1.** *Espérance de l'indicatrice d'un événement.* L'indicatrice $\mathbf{1}_A$ de l'événement A est égale à 1 si A est réalisé et à 0 dans le cas contraire. Soit p la

probabilité de A. On calcule alors l'espérance mathématique :

$$\mathrm{E}\{\mathbf{1}_A\} = 1 \times p + 0 \times (1-p) = p\,. \tag{5.3}$$

Quelques propriétés.

Nous donnons, à titre d'exemple, quelques propriétés concernant la linéarité de l'espérance mathématique des variables aléatoires discrètes.

- Si c est une constante, $\mathrm{E}\{X = c\} = c$.
- Si $\mathrm{E}\{X\}$ et $\mathrm{E}\{Y\}$ existent, on a $\mathrm{E}\{cX\} = c\,\mathrm{E}\{X\}$ et $\mathrm{E}\{X+Y\} = \mathrm{E}\{X\} + \mathrm{E}\{Y\}$.

On donnera plus loin, au chapitre 5.1.4, un tableau plus complet de ces propriétés lorsqu'on aura défini l'espérance mathématique d'un variable aléatoire quelconque (discrète ou continue).

Espérance mathématique conditionnelle.

L'espérance mathématique *conditionnelle* de la variable aléatoire X vis-à-vis de l'événement A est définie par :

$$\mathrm{E}\{X|A\} = \sum_i x_i \ \Pr\{X = x_i|A\}\,, \tag{5.4}$$

où la somme s'étend à tous les indices numérotant les valeurs possibles de X. D'après cette définition l'espérance mathématique conditionnelle de X n'est autre que l'espérance mathématique de la loi conditionnelle de X. Cette remarque nous permet de généraliser le théorème des probabilités totales en un théorème des *espérances mathématiques totales.*

Théorème 5.1. *Si $\{A_k\}$, $k \in \mathbb{N}$ désigne un système complet d'événements disjoints et si X est une variable aléatoire discrète dont l'espérance mathématique* $\mathrm{E}\{X\}$ *existe, on a :*

$$\mathrm{E}\{X\} = \sum_k \mathrm{E}\{X|A_k\}\,\Pr\{A_k\} \tag{5.5}$$

La démonstration se fait à l'aide du théorème des probabilités totales. Soient x_i les valeurs prises par X. Il vient :

$$\mathrm{E}\{X\} = \sum_i x_i \Pr\{X = x_i\} = \sum_i x_i \sum_k \Pr\{X = x_i|A_k\}\,\Pr\{A_k\}\,. \tag{5.6}$$

Comme $\mathrm{E}\{X\}$ existe, on peut réarranger l'ordre des sommations:

$$\mathrm{E}\{X\} = \sum_k \sum_i x_i \Pr\{X = x_i|A_k\}\,\Pr\{A_k\}\,, \tag{5.7}$$

ce qui, d'après (5.4) s'écrit :

$$\mathrm{E}\{X\} = \sum_k \mathrm{E}\{X|A_k\}\,\Pr\{A_k\}\,, \tag{5.8}$$

On vérifie facilement que ce théorème recouvre aussi celui des probabilités totales en prenant pour variable aléatoire l'indicatrice d'un certain événement B et en sachant que $\mathrm{E}\{\mathbf{1}_B\} = \Pr\{B\}$ et $\mathrm{E}\{\mathbf{1}_B|A_k\} = \Pr\{B|A_k\}$.

Espérance d'une fonction de la variable aléatoire discrète.

Soit Y une variable aléatoire définie par le changement de variable $Y = \varphi(X)$. Par définition l'espérance mathématique de Y est, si elle existe, égale à $\mathrm{E}\{Y\} = \sum_j y_j \Pr\{Y = y_j\}$. Si on désigne par $x_j^{(k)}$, $k = 1, 2, \ldots$ les solutions de l'équation $y_j = \varphi(x)$, on a :

$$\begin{aligned}\mathrm{E}\{Y\} &\equiv \sum_j y_j \Pr\{Y = y_j\} = \sum_j y_j \sum_k \Pr\{X = x_j^{(k)}\} \\ &= \sum_{jk} y_j \Pr\{X = x_j^{(k)}\} = \sum_{jk} \varphi(x_j^{(k)}) \Pr\{X = x_j^{(k)}\} .\end{aligned}$$

Le changement de variable étant défini pour toutes les valeurs x_i de X, la somme double recouvre tous les x_i. De plus φ est une fonction et à un $x_j^{(k)}$ ne correspond qu'un seul y_j, la somme double ne « compte » pas deux fois le même x_i. En conclusion la somme double est identique à une somme sur tous les x_i. Il vient alors :

$$\sum_{jk} \varphi(x_j^{(k)}) \Pr\{X = x_j^{(k)}\} = \sum_i \varphi(x_i) \Pr\{X = x_i\} .$$

On peut alors écrire :

$$\mathrm{E}\{Y\} = \mathrm{E}\{\varphi(X)\} \quad \text{pour} \quad Y = \varphi(X) . \tag{5.9}$$

Ainsi, pour calculer l'espérance de $Y = \varphi(X)$, il n'est pas nécessaire de calculer la loi suivie par Y (c'est-à-dire $\Pr\{Y = y_i\}$), il suffit de calculer l'espérance de $\varphi(X)$. Ce résultat très important se généralise aux variables aléatoires quelconques.

5.1.2 L'espérance mathématique des variables aléatoires continues.

La généralisation de la notion d'espérance mathématique aux variables aléatoires continues passe par l'intermédiaire de la construction d'une variable aléatoire discrète associée, puis par passage à la limite. Ce passage à la limite conserve pour les variables aléatoires continues les principales propriétés des variables aléatoires discrètes.

Variable aléatoire discrète associée.

Soit X une variable aléatoire de fonction de répartition F dont les valeurs possibles appartiennent, par exemple, à tout l'axe réel $\mathbb{R}$. Limitons-nous à une description plus grossière des valeurs possibles de X en découpant l'axe réel en cellules identiques de dimension h ($h > 0$). Acceptons ensuite une certaine perte de précision dans la description des issues de X en attribuant la même valeur à deux issues différentes mais appartenant à la même cellule. Notons cependant que c'est bien ce à quoi il faut se résoudre au cours d'une expérience physique réelle. On dit alors que l'on a décrit les issues de X avec la résolution h.

Supposons, pour simplifier, que nous avons découpé l'axe réel à partir de l'origine de telle sorte que la cellule numéro k : Δ_k contienne toutes les issues

de X comprises entre kh et $(k+1)h$. Plus précisément, dans cette cellule Δ_k la variable aléatoire X est telle que $kh \leq X < (k+1)h$; nous lui attribuerons alors la valeur kh. La probabilité $p_k = \Pr\{X \in \Delta_k\}$ pour que X appartienne à cet intervalle est telle que :

$$p_k = F((k+1)h) - F(kh) + \Pr\{X = kh\} - \Pr\{X = (k+1)h\}\,. \tag{5.10}$$

Si, comme nous l'avons supposé, la variable X est continue, alors les deux dernières probabilités sont nulles.

Les issues de X ainsi décrites sont les mêmes que celles d'une variable aléatoire discrète X_h égale au plus grand multiple de h inférieur ou égal à X et prenant la valeur kh avec la probabilité p_k, soit :

$$X_h = h\left\lfloor \frac{X}{h} \right\rfloor, \quad \Pr\{X_h = kh\} = p_k\,. \tag{5.11}$$

L'espérance mathématique de X_k est par définition égale à :

$$\mathrm{E}\{X_h\} = \sum_{k=-\infty}^{+\infty} kh \Pr\{X_h = kh\} = \sum_{k=-\infty}^{+\infty} kh \Pr\{kh \leq X < (k+1)h\}\,. \tag{5.12}$$

Nous définirons l'espérance mathématique $\mathrm{E}\{X\}$ d'une variable aléatoire continue X comme la limite, si elle existe, de $\mathrm{E}\{X_k\}$ quand $h \to 0$, soit :

$$\mathrm{E}\{X\} = \lim_{h\to 0} \sum_{k=-\infty}^{+\infty} kh \Pr\{kh \leq X < (k+1)h\}\,. \tag{5.13}$$

Cette définition coïncide avec l'intégrale de Lebesgue de la fonction $X(\omega)$ pondérée par la mesure Pr. Comme pour les variables aléatoires discrètes, il faut que l'espérance mathématique converge vers une valeur qui soit indépendante de l'ordre dans lequel on effectue la somme, ce qui impose que l'intégrale soit absolument convergente. On a alors :

$$\mathrm{E}\{X\} = \int_\Omega X(\omega)\, dP \quad \text{pour} \quad \int_\Omega |X(\omega)|\, dP < \infty\,. \tag{5.14}$$

Dans le cas des variables aléatoires continues l'équation (5.10) nous montre que la mesure dP est égale à dF et l'intégrale de Lebesgue se réduit à l'intégrale de Stieltjes suivante :

$$\mathrm{E}\,\{X\} = \int_{-\infty}^{\infty} x\, dF \quad \text{pour} \quad \int_{-\infty}^{\infty} |x|\, dF < \infty\,. \tag{5.15}$$

5.1.3 L'espérance mathématique des variables aléatoires quelconques.

Si la variable aléatoire X est mixte, c'est-à-dire si sa fonction de répartition F présente des discontinuités, son espérance mathématique est toujours égale à l'intégrale de Stieltjes car celle-ci prend en compte les discontinuités de F. Cette formulation inclut aussi le cas des variables aléatoires discrètes car dans ce cas

F est une fonction « en escalier » présentant les sauts $p_i = F^+(x_i) - F^-(x_i)$ aux points de discontinuité x_i et l'intégrale de Stieltjes se réduit alors à une somme :

$$\mathrm{E}\{X\} = \sum_i x_i\, p_i \ . \tag{5.16}$$

Si F admet une densité f, l'intégrale de Stieltjes se réduit à l'intégrale de Riemann classique :

$$\mathrm{E}\{X\} = \int_{-\infty}^{\infty} x f(t)\, dt \quad \text{pour} \quad \int_{-\infty}^{\infty} |x| f(t)\, dt < \infty \ . \tag{5.17}$$

D'un point de vue théorique, l'espérance mathématique est une fonctionnelle, c'est-à-dire une fonction dont l'argument est une fonction et dont le résultat est un scalaire. La fonction sur laquelle agit l'espérance mathématique est la variable aléatoire $X = X(\omega)$. En physique, l'espérance mathématique est appelée « *moyenne d'ensemble* » et est souvent notée $< X >$.

5.1.4 Propriétés de l'espérance mathématique.

Les propriétés de l'espérance mathématique sont celles qui découlent des propriétés de l'intégrale de Lebesgue. Nous en donnons ci-dessous un certain nombre, les plus remarquables concernent la linéarité de l'espérance mathématique.

1. Si c est une constante, $\mathrm{E}\{X = c\} = c$. En particulier $\mathrm{E}\{\mathrm{E}\{X\}\} = \mathrm{E}\{X\}$.

2. On a $\mathrm{E}\{cX\} = c\,\mathrm{E}\{X\}$.

3. Si $\mathrm{E}\{X\}$ et $\mathrm{E}\{Y\}$ existent, alors $\mathrm{E}\{X+Y\} = \mathrm{E}\{X\}+\mathrm{E}\{Y\}$. En particulier $\mathrm{E}\{X - \mathrm{E}\{X\}\} = 0$

4. Plus généralement $\mathrm{E}\{X_1+X_2+\cdots+X_n\} = \mathrm{E}\{X_1\}+\mathrm{E}\{X_2\}+\cdots+\mathrm{E}\{X_n\}$.

5. Plus généralement encore l'espérance mathématique est une fonctionnelle linéaire. Si les c_i sont des constantes et si l'espérance des variables aléatoires X_i existe, alors l'espérance d'une combinaison linéaire de ces variables aléatoires existe également et on a :

$$\mathrm{E}\Big\{\sum_i c_i X_i\Big\} = \sum_i c_i\, \mathrm{E}\{X_i\} \, . \tag{5.18}$$

6. L'espérance mathématique du produit de deux variables aléatoires quelconques n'est en général pas égale au produit de leur espérance mathématique ($\mathrm{E}\{XY\} \neq \mathrm{E}\{X\}\,\mathrm{E}\{Y\}$). Mais si $\mathrm{E}\{X^2\}$ et $\mathrm{E}\{Y^2\}$ existent, on a l'inégalité de Cauchy-Schwarz :

$$|\mathrm{E}\{XY\}| \leq \sqrt{\mathrm{E}\{X^2\}\,\mathrm{E}\{Y^2\}} \, . \tag{5.19}$$

7. En revanche, si X et Y sont des variables aléatoires *indépendantes* alors $\mathrm{E}\{XY\} = \mathrm{E}\{X\}\,\mathrm{E}\{Y\}$.

8. Si X est une variable aléatoire positive ou nulle alors son espérance est aussi positive ou nulle. L'espérance n'est nulle que si X est elle-même une variable aléatoire identiquement nulle :

$$X \geq 0 \Rightarrow \mathrm{E}\{X\} \geq 0\,, \tag{5.20}$$

$$X = 0 \iff \mathrm{E}\{X\} = 0\,. \tag{5.21}$$

5.1.5 Espérance mathématique conditionnelle.

La fonction de répartition conditionnelle permet de calculer des espérances mathématiques conditionnelles :

$$\mathrm{E}\{X|A\} = \int_{-\infty}^{\infty} x\, dF_{X|A}\,. \tag{5.22}$$

Espérances mathématiques conditionnelles totales.

Si les A_i représentent les événements d'un système complet $\mathcal{A}$ recouvrant un sous-ensemble Ω', on tire facilement de (2.44) :

$$\mathrm{E}\{X|\Omega'\} = \sum_i \mathrm{E}\{X|A_i\}\,\mathrm{Pr}\{A_i|\Omega'\}\,, \tag{5.23}$$

et de (2.45) dans le cas où Ω' est confondu avec l'espace des épreuves Ω :

$$\mathrm{E}\{X\} = \sum_i \mathrm{E}\{X|A_i\}\,\mathrm{Pr}\{A_i\}\,. \tag{5.24}$$

Ces équations peuvent être considérées comme les espérances des variables aléatoires discrètes $\mathrm{E}\{X|A_i\}$. Cette remarque nous conduit à réécrire les formules (5.23) et (5.24) sous la forme :

$$\mathrm{E}\{X|\Omega'\} = \mathrm{E}\{\mathrm{E}\{X|\mathcal{A}\}|\Omega'\}\,, \tag{5.25}$$

et si Ω' est tout l'ensemble Ω :

$$\mathrm{E}\{X\} = \mathrm{E}\{\mathrm{E}\{X|\mathcal{A}\}\}\,. \tag{5.26}$$

L'analogie mécanique des formules précédentes exprime simplement le fait que le centre de gravité d'une distribution de masse quelconque peut se calculer d'abord sur des éléments disjoints de cette distribution, puis comme barycentre des centres de gravité des éléments, les centres de gravité des éléments disjoints étant pondérés par la masse de l'élément correspondant.

▶**Exemple 5.2.** Une variable aléatoire X est la somme de N variables aléatoires X_n, $X = \sum_{n=0}^{N} X_n$. Le nombre N est une variable aléatoire discrète indépendante des X_n et $\mathrm{Pr}\{N = n\} = p_n$. Les X_n possèdent tous la même espérance $\mathrm{E}\{X_n\} = \mu_X$ et on a $\mathrm{E}\{N\} = \mu_N$. On demande l'espérance de X.

Les événements $A_n = \{N = n\}$ formant un système complet d'événements indépendants on peut appliquer les formules (5.24) ou (5.26). Il vient :

$$\mathrm{E}\{X|N = n\} = \mathrm{E}\{\sum_{i=0}^{n} X_i\} = \sum_{i=0}^{n} \mathrm{E}\{X_i\} = n\mu_X\,,$$

$$\mathrm{E}\{X\} = \sum_{n=0}^{\infty} n\mu_X p_n = \mu_X \sum_{i=0}^{\infty} n p_n = \mu_X \mu_N\,.$$

5.1.6 Espérance d'une fonction de la variable aléatoire.

Comme dans le cas discret, l'espérance mathématique d'une fonction $Y = \varphi(X)$ de la variable aléatoire X est donnée par l'expression :

$$\mathrm{E}\{Y\} = \mathrm{E}\{\varphi(X)\}\,. \tag{5.27}$$

Là non plus il n'est pas nécessaire de calculer la fonction de répartition de Y afin d'évaluer son espérance. Une égalité du type (5.27) signifie que la fonctionnelle $\mathrm{E}\{\ \}$ peut être envisagée comme l'espérance de Y pour la loi suivie par $\varphi(X)$ ou comme l'espérance de $\varphi(X)$ pour la loi suivie par X. L'identité entre ces deux interprétations vient de ce que l'on a imposé l'absolue convergence de l'intégrale définissant l'espérance. Comme nous l'avons vu pour les variables aléatoires discrètes, c'est à cette condition que l'espérance reste définie quel que soit l'ordre dans lequel on effectue les sommations.

▶**Exemple 5.3.** *Espérance du carré d'une variable aléatoire.* Si une variable aléatoire X possède une moyenne μ et une variance σ^2, d'après (2.29) l'espérance mathématique de X^2 existe et est égale à $\mu^2 + \sigma^2$. On a :

$$\mathrm{E}\{X^2\} = \mu^2 + \sigma^2 \quad \text{avec} \quad \mu = \mathrm{E}\{X\}, \quad \sigma^2 = \mathrm{E}\{(X - \mathrm{E}\{X\})^2\}\,. \tag{5.28}$$

Si X possède un moment d'ordre 4 : μ_4, X^2 possède alors une variance :

$$\mathrm{Var}(X^2) = \mu_4 + 4\mu_3\mu + 4\sigma^2\mu^2 - \sigma^4 \tag{5.29}$$

5.2 Inégalités impliquant des espérances.

Nous n'établissons ici que les inégalités du type Cauchy-Schwarz, on trouvera d'autres inégalités dans Loève [49] chap. II sec. 7 et dans l'ouvrage "inequalities" de Hardy, Littlewood et Pòlya [29].

5.2.1 L'inégalité de Cauchy-Schwarz.

L'inégalité de Cauchy-Schwarz porte sur les valeurs relatives des variances et de la covariance de deux variables aléatoires (X_1, X_2), elle fournit en outre une mesure du degré de dépendance linéaire entre ces deux variables.

Théorème 5.2. *Si X_1 et X_2 sont des variables aléatoires quelconques et si $\mathrm{E}\{X_1^2\}$ et $\mathrm{E}\{X_2^2\}$ existent, alors $\mathrm{E}\{X_1X_2\}$ existe aussi et on a :*

$$\mathrm{E}\{X_1X_2\}^2 \leq \mathrm{E}\{X_1^2\}\,\mathrm{E}\{X_2^2\}\,, \tag{5.30}$$

l'égalité n'ayant lieu que si, et seulement si, les variables aléatoires X_1 et X_2 sont linéairement dépendantes, c'est-à-dire s'il existe deux nombres λ_1 et λ_2 non tous nuls tels que $\lambda_1X_1 + \lambda_2X_2 = 0$, presque partout. Dans ce cas on a :

$$\lambda_1\,\mathrm{E}\{X_1^2\} + \lambda_2\,\mathrm{E}\{X_1X_2\} = 0, \ \textit{si}\ \lambda_2 \neq 0 \tag{5.31}$$

$$\lambda_2\,\mathrm{E}\{X_2^2\} + \lambda_1\,\mathrm{E}\{X_1X_2\} = 0, \ \textit{si}\ \lambda_1 \neq 0\,. \tag{5.32}$$

Démonstration. Considérons la variable aléatoire $Y = \lambda_1X_1 + \lambda_2X_2$, où λ_1, λ_2 sont deux nombres quelconques non tous nuls. Introduisons de plus la forme quadratique $Q(\lambda_1, \lambda_2) = \mathrm{E}\{(\lambda_1X_1 + \lambda_2X_2)^2\} \geq 0$, l'égalité n'ayant lieu que si $\lambda_1X_1 + \lambda_2X_2 = 0$

(presque partout). Nous supposerons que λ_2 est différent de zéro. Si tel n'était pas le cas nous échangerions les rôles joués par λ_1 et λ_2 dans la démonstration.

Envisageons tout d'abord le cas $Y \neq 0$: en développant $\mathrm{E}\{Y^2\}$, on obtient $\lambda_1^2\,\mathrm{E}\{X_1^2\} + 2\lambda_1\lambda_2\,\mathrm{E}\{X_1X_2\} + \lambda_2^2\,\mathrm{E}\{X_2^2\} > 0$. C'est un trinome du second degré en λ_1 qui n'est positif que si son discriminant est négatif, c'est-à-dire si $\lambda_2^2(\mathrm{E}\{X_1X_2\}^2 - \mathrm{E}\{X_1^2\}\,\mathrm{E}\{X_2^2\}) < 0$. Comme $\lambda_2 \neq 0$, il vient $\mathrm{E}\{X_1X_2\}^2 - \mathrm{E}\{X_1^2\}\,\mathrm{E}\{X_2^2\} < 0$, ce qui démontre la première partie du théorème.

Il faut maintenant envisager le cas où les X_1, X_2 sont linéairement dépendants, c'est-à-dire $Y = 0$. Comme nous avons supposé $\lambda_2 \neq 0$, on peut écrire $X_2 = \mu X_1$ ($\mu = -\lambda_1/\lambda_2$). Il vient $\mathrm{E}\{X_2^2\} = \mu^2\,\mathrm{E}\{X_1^2\}$, $\mathrm{E}\{X_1X_2\} = \mu\,\mathrm{E}\{X_1^2\}$ et $\mathrm{E}\{X_1X_2\}^2 - \mathrm{E}\{X_1^2\}\,\mathrm{E}\{X_2^2\} = 0$. Réciproquement si $\mathrm{E}\{X_1X_2\}^2 - \mathrm{E}\{X_1^2\}\,\mathrm{E}\{X_2^2\} = 0$, cela veut dire que la forme Q s'annule pour la racine (double) $\lambda_1 = -\lambda_2\,\mathrm{E}\{X_1X_2\}/\mathrm{E}\{X_1^2\}$ ce qui n'est possible que si $Y = \lambda_1X_1 + \lambda_2X_2 = 0$ (presque partout). Ceci établit la dépendance linéaire de X_1 et X_2 pour ce λ_1. Dans le cas où $\lambda_1 \neq 0$ on trouverait une condition sur λ_2 qui établirait (5.32). □

Si l'on applique l'inégalité de Cauchy-Schwarz pour les variables centrées $X_1 - \mu_1$ et $X_2 - \mu_2$ ($\mu_1 = \mathrm{E}\{X_1\}, \mu_2 = \mathrm{E}\{X_2\}$) et à la condition que X_1, X_2 possèdent des variances σ_1^2 et σ_2^2 on obtient $\rho^2\sigma_1^2\sigma_2^2 \leq \sigma_1^2\sigma_2^2$ où ρ est le coefficient de corrélation, $\rho = \mathrm{E}\{(X_1 - \mu_1)(X_2 - \mu_2)\}/\sigma_1\sigma_2$. Cela implique que $|\rho| = 1$ et que, d'après (5.31) ou (5.32), il existe une relation affine entre X_1 et X_2 :

$$|\rho| = 1 \Leftrightarrow \frac{X_2 - \mu_2}{\sigma_2} = \rho\frac{X_1 - \mu_1}{\sigma_1}\,. \tag{5.33}$$

Cela justifie l'utilisation du coefficient de corrélation comme mesure de la dépendance linéaire des variables $X_1 - \mu_1$ et $X_2 - \mu_2$.

5.2.2 Les inégalités de Cauchy-Schwarz d'ordre n.

L'inégalité de Cauchy-Schwarz est sujette à généralisation en considérant la variable $Y = \lambda_1X_1 + \lambda_2X_2 + \cdots + \lambda_nX_n$. On dispose alors du théorème suivant :

Théorème 5.3. *Si les variables aléatoires sont quelconques mais possèdent toutes des moments d'ordre 2: $\mathrm{E}\{X_iX_j\} < \infty$, alors la matrice $\boldsymbol{R}$ des moments d'ordre 2 est définie non-négative. La matrice $\boldsymbol{R}$ est définie positive si, et seulement si, les X_i sont linéairement indépendants.*

Avant d'entreprendre la démonstration, donnons un exemple pour trois variables aléatoires X_1, X_2 et X_3. Leur matrice $\boldsymbol{R}$ est par définition égale à :

$$\boldsymbol{R} = \begin{pmatrix} \mathrm{E}\{X_1^2\} & \mathrm{E}\{X_1X_2\} & \mathrm{E}\{X_1X_3\} \\ \mathrm{E}\{X_2X_1\} & \mathrm{E}\{X_2^2\} & \mathrm{E}\{X_2X_3\} \\ \mathrm{E}\{X_3X_1\} & \mathrm{E}\{X_3X_2\} & \mathrm{E}\{X_3^2\} \end{pmatrix}.$$

Plaçons-nous dans le cas de l'indépendance linéaire des X_1, X_2, X_3. L'inégalité de Cauchy-Schwarz d'ordre 3 nous dit que $\boldsymbol{R}$ est définie positive, ce qui implique, et réciproquement, que ses mineurs principaux sont positifs. Exprimons les. Il vient :

$$\mathrm{E}\{X_1^2\} > 0,\ \begin{vmatrix} \mathrm{E}\{X_1^2\} & \mathrm{E}\{X_1X_2\} \\ \mathrm{E}\{X_2X_1\} & \mathrm{E}\{X_2^2\} \end{vmatrix} > 0,\ \begin{vmatrix} \mathrm{E}\{X_1^2\} & \mathrm{E}\{X_1X_2\} & \mathrm{E}\{X_1X_3\} \\ \mathrm{E}\{X_2X_1\} & \mathrm{E}\{X_2^2\} & \mathrm{E}\{X_2X_3\} \\ \mathrm{E}\{X_3X_1\} & \mathrm{E}\{X_3X_2\} & \mathrm{E}\{X_3^2\} \end{vmatrix} > 0\,.$$

La première inégalité est triviale, la deuxième est l'inégalité de Cauchy-Schwarz classique et la troisième est l'inégalité de Cauchy-Schwarz d'ordre 3. En introduisant les coefficients de corrélation ρ_{ij} entre les variables X_i, X_j on obtient :

$$\begin{vmatrix} 1 & \rho_{12} \\ \rho_{12} & 1 \end{vmatrix} > 0, \quad \begin{vmatrix} 1 & \rho_{12} & \rho_{13} \\ \rho_{12} & 1 & \rho_{23} \\ \rho_{13} & \rho_{23} & 1 \end{vmatrix} > 0, \tag{5.34}$$

et des inégalités similaires en permutant les indices.

Démonstration. On pose $Y = \sum_{i=1}^{n} \lambda_i X_i$ et $Q(\lambda_1, \ldots, \lambda_n) = \mathrm{E}\{Y^2\} \geq 0$, l'égalité n'étant assurée que si $Y = 0$ presque partout. Supposons $Y \neq 0$: dans ce cas Q, qui est une forme homogène de degré deux, est définie positive : $Q = \sum_{i,j=1}^{n} \lambda_i \lambda_j \, \mathrm{E}\{X_i X_j\} > 0$. Sa matrice caractéristique dont les termes valent $\frac{1}{2}\partial^2 Q / \partial\lambda_i \partial\lambda_j$ est également définie positive. On a $\frac{1}{2}\partial^2 Q / \partial\lambda_i \partial\lambda_j = \mathrm{E}\{X_i X_j\}$, ce qui établit le théorème direct.

Réciproquement si la matrice caractéristique est définie positive, Q est positive, ce qui implique $\mathrm{E}\{(\sum \lambda_i X_i)^2\} > 0$ ce qui est impossible si les X_i sont linéairement dépendants. Les X_i sont donc linéairement indépendants. □

5.3 Nombres caractéristiques.

5.3.1 Les moments.

L'espérance mathématique nous permet de redéfinir les moments des lois de probabilité. Soient μ'_r et μ_r les moments respectivement non-centrés et centrés d'ordre r. On a :

$$\mu'_r = \mathrm{E}\{X^r\} \quad \text{et} \quad \mu_r = \mathrm{E}\{(X - \mathrm{E}\{X\})^r\}, \quad r = 1, 2, \ldots \tag{5.35}$$

Pour qu'un moment centré ou non-centré d'ordre r existe il faut et il suffit que $\mathrm{E}\{|X|^r\}$ existe, soit :

$$\int_{-\infty}^{\infty} |x|^r \, dF < \infty \, .$$

Si le moment d'ordre r existe alors tous les autres moments d'ordres inférieurs à r existent aussi.

5.3.2 L'erreur quadratique moyenne.

Cette quantité fait référence à une valeur particulière a de $\mathbb{R}$. Elle est définie par l'expression :

$$q_X^2(a) = \mathrm{E}\left\{(X - a)^2\right\} \, . \tag{5.36}$$

Lorsque le point a est pris égal à la moyenne $\mu = \mathrm{E}\{X\}$ de la densité, l'erreur quadratique moyenne de X est égale à la variance de X. L'erreur quadratique moyenne de X calculée autour d'une valeur a différente de la moyenne de X est toujours supérieure à la variance de X. En effet :

$$\begin{aligned} \mathrm{E}\left\{(X - a)^2\right\} &= \mathrm{E}\left\{(X - \mu + \mu - a)^2\right\} \\ &= \mathrm{E}\left\{(X - \mu)^2\right\} + 2\,\mathrm{E}\left\{(X - \mu)(\mu - a)\right\} + \mathrm{E}\left\{(\mu - a)^2\right\} \\ &= \mathrm{E}\left\{(X - \mu)^2\right\} + 2(\mu - \mu)(\mu - a) + (\mu - a)^2 \\ &= \mathrm{Var}(X) + (\mu - a)^2 \end{aligned}$$

La quantité $\mu - a$ est appelée le *biais*. Le carré du biais : $(\mu - a)^2$ est toujours strictement positif si $\mu \neq a$, il n'est nul que si $\mu = a$. L'erreur quadratique moyenne est donc bien minimum lorsqu'elle est calculée autour de la moyenne de la variable aléatoire X. Ces propriétés analogues au théorème de Huygens de la mécanique sont contenues dans la relation : « Erreur quadratique moyenne = variance + carré du biais » :

$$q_X^2(a) = \mathrm{E}\left\{(X-a)^2\right\} = \sigma^2 + (\mu - a)^2 . \tag{5.37}$$

C'est l'équation d'une parabole dont le minimum est en μ. L'expression (5.37), analogue au théorème de Pythagore, est illustrée par la figure 5.1.

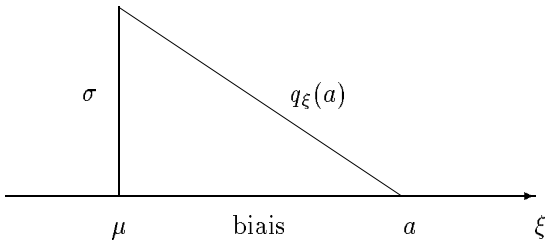


FIG. 5.1: *Relation « de Pythagore » reliant la variance, le biais et l'erreur quadratique moyenne de la variable aléatoire ξ calculée autour de la valeur a.*

5.4 Fonctions caractéristiques

L'espérance mathématique permet de redéfinir la fonction de répartition, la densité de probabilité et d'introduire deux nouvelles fonctions importantes.

5.4.1 La fonction de répartition.

La fonction de répartition est l'espérance de la fonction indicatrice suivante :

$$F(x) = \mathrm{E}\left\{\mathbf{1}_{]-\infty,x]}\right\} = \int_{-\infty}^{x} dF . \tag{5.38}$$

5.4.2 La densité de probabilité.

On trouve la densité de probabilité par dérivation et donc comme espérance des translatées de δ :

$$f(x) = \mathrm{E}\left\{\delta_x\right\} = \int_{-\infty}^{\infty} \delta(u-x) f(u)\, du . \tag{5.39}$$

5.4.3 La fonction caractéristique.

La fonction caractéristique Z est définie comme l'espérance des complexes de module unité :

$$Z(\omega) = \mathrm{E}\{e^{iX\omega}\} = \int_{-\infty}^{\infty} e^{iu\omega}\, dF(u) . \tag{5.40}$$

Si la densité de probabilité existe, on a :

$$Z(\omega) = \int_{-\infty}^{\infty} e^{iu\omega} f(u)\, du \ . \tag{5.41}$$

La fonction caractéristique est alors la transformée de Fourier de la densité de probabilité. Inversement on trouve la densité de probabilité à partir de la fonction caractéristique grâce à la formule réciproque :

$$\tfrac{1}{2}(f(x^+) + f(x^-)) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-iux} Z(u)\, du \ . \tag{5.42}$$

La fonction caractéristique possède les propriétés suivantes :

1. $Z(0) = 1$, $|Z(\omega)| \leq 1$, pour $\omega \in]-\infty, +\infty[$.
2. Si le moment d'ordre n existe, on a $\left.\dfrac{d^n Z(\omega)}{d\omega^n}\right|_{\omega=0} = i^n \, \mathrm{E}\{X^n\}$.
3. La fonction caractéristique est hermitienne : $\overline{Z(-\omega)} = Z(\omega)$. C'est-à-dire que sa partie réelle est symétrique alors que sa partie imaginaire est anti-symétrique.
4. La fonction caractéristique d'une somme de variables aléatoires indépendantes est égale au produit de leurs fonctions caractéristiques : $Z_{X_1+X_2}(\omega) = Z_{X_1}(\omega) Z_{X_2}(\omega)$.

5.4.4 La fonction génératrice des moments.

La fonction génératrice des moments M est définie par la transformée de Hilbert de la densité de probabilité :

$$M(x) = \mathrm{E}\left\{(1 - Xx)^{-1}\right\} = \int_{-\infty}^{\infty} (1 - ux)^{-1} f(u)\, du \ . \tag{5.43}$$

La fonction M tire son nom de la propriété suivante :

$$M(x) = \sum_{k=0}^{\infty} x^k \, \mathrm{E}\left\{X^k\right\} ; \quad |Xx| < 1 \ . \tag{5.44}$$

5.5 Espérance des variables aléatoires d'un couple.

L'espérance mathématique de chacune des variables aléatoires X et Y d'un couple (X, Y) de fonction de répartition F est par définition calculée à l'aide des fonctions de répartition marginales de F :

$$\mathrm{E}\{X\} = \int_{-\infty}^{\infty} x\, dF_X \ , \quad \mathrm{E}\{Y\} = \int_{-\infty}^{\infty} y\, dF_Y \ , \tag{5.45}$$

où F_X det F_Y désignent respectivement la fonction de répartition de X et la fonction de répartition de Y. Supposons pour simplifier que la loi F possède une

densité f et ne considérons que la variable X. On a alors par définition des lois marginales (voir l'équation (3.11)) :

$$\int_{-\infty}^{\infty} x\, dF_X = \int_{-\infty}^{\infty} x\, dx \int_{-\infty}^{\infty} f(x,y)\, dy = \iint_{\mathbb{R}^2} x f(x,y)\, dxdy\,,$$

soit :

$$\mathrm{E}\{X\} = \iint_{\mathbb{R}^2} x f(x,y)\, dxdy\,, \quad \mathrm{E}\{Y\} = \iint_{\mathbb{R}^2} y f(x,y)\, dxdy\,. \tag{5.46}$$

L'espérance de X peut donc être calculée soit avec sa propre loi, soit avec la loi conjointe de X et d'une autre variable aléatoire, ou plus généralement de plusieurs autres variables aléatoires.

De même pour une fonction φ des variables aléatoires on a :

$$\mathrm{E}\{\varphi(X,Y)\} = \iint_{\mathbb{R}^2} \varphi(x,y)\, dF\,, \tag{5.47}$$

et si la loi possède une densité :

$$\mathrm{E}\{\varphi(X,Y)\} = \iint_{\mathbb{R}^2} \varphi(x,y) f(x,y)\, dxdy\,. \tag{5.48}$$

5.5.1 Espérances conditionnelles des lois 2D.

L'espérance mathématique conditionnelle est définie comme l'espérance des lois conditionnelles, ainsi :

$$\mathrm{E}\{\varphi(X,Y)|\Omega'\} = \iint_{\mathbb{R}^2} \varphi(u,v) f(u,v|\Omega')\, dxdy\,. \tag{5.49}$$

5.5.2 Espérance des lois nD.

Les remarques faites dans le cas 2D se généralisent naturellement au cas nD ; par exemple si la variable aléatoire Y est fonction de plusieurs variables aléatoires X_i, $Y = \varphi(X_1, \ldots, X_n)$, on peut montrer que :

$$\mathrm{E}\{Y\} = \int_{\mathbb{R}^n} \cdots \int \varphi(x_1, \ldots, x_n) f(x_1, \ldots, x_n)\, dx_1 \ldots dx_n\,. \tag{5.50}$$

Soit encore :

$$\mathrm{E}\{Y\} = \mathrm{E}\{\varphi(X_1, \ldots, X_n)\}\,. \tag{5.51}$$

5.5.3 Espérance mathématique d'une matrice.

L'espérance mathématique d'une matrice $\boldsymbol{A}$ dont les éléments sont des variables aléatoires est, par définition, égale à une matrice dont les éléments sont les

espérances des éléments de $\boldsymbol{A}$. Avec cette définition la matrice $\boldsymbol{V}$ des variances-covariances s'écrit alors :

$$\boldsymbol{V} = \mathrm{E}\left\{(\boldsymbol{X} - \mathrm{E}\{\boldsymbol{X}\})(\boldsymbol{X} - \mathrm{E}\{\boldsymbol{X}\})^t\right\} . \tag{5.52}$$

On établit immédiatement l'identité :

$$\boldsymbol{V} = \mathrm{E}\{\boldsymbol{X}\,\boldsymbol{X}^t\} - \mathrm{E}\{\boldsymbol{X}\}\,\mathrm{E}\{\boldsymbol{X}\}^t \tag{5.53}$$

5.6 Caractéristiques numériques des fonctions de variables aléatoires.

Nous avons vu que la densité de probabilité du quotient de deux variables aléatoires normales menait, dans ce cas pourtant simple, à une expression déjà relativement compliquée. Dans la pratique, il devient vite impossible de calculer avec exactitude la densité de probabilité d'une fonction quelconque de variables aléatoires, et il faut alors se limiter au calcul de certaines de ses caractéristiques numériques comme sa moyenne et son écart type, s'ils existent. Bien souvent même, il faudra se contenter de valeurs approximatives ou asymptotiques.

5.6.1 Quantiles d'une fonction de la variable aléatoire.

On trouve facilement l'expression des quantiles de la loi suivie par Y connaissant ceux de la loi F suivie par X lorsque le changement de variable aléatoire $Y = \varphi(X)$ est bijectif. Soit x_α un quantile de la loi F. On a par définition du quantile $F(x_\alpha) = 1 - \alpha$ et par définition de la fonction de répartition $\Pr\{X \leq x_\alpha\} = 1 - \alpha$. Supposons que φ est une fonction croissante. On a $\Pr\{X \leq x_\alpha\} = \Pr\{Y \leq \varphi(x_\alpha)\}$, ce qui montre que le quantile d'ordre α de la loi suivie par Y est égal à $\varphi(x_\alpha)$. Pour une fonction décroissante et pour des variables aléatoires strictement continues, on trouve que le quantile d'ordre α de Y est égal à $\varphi(x_{1-\alpha})$. Le tableau 5.1 résume ces résultats.

Variable aléatoire	X	$\varphi(X)$, $\varphi' > 0$	$\varphi(X)$, $\varphi' < 0$
Quantile d'ordre α	x_α	$\varphi(x_\alpha)$	$\varphi(x_{1-\alpha})$
Médiane	$x_{0.5}$	$\varphi(x_{0.5})$	$\varphi(x_{0.5})$

TAB. 5.1: *Quantile de la loi suivie par la variable aléatoire $\varphi(X)$ lorsque X est une variable aléatoire continue et que le changement de variable φ est bijectif.*

5.6.2 Moments d'une fonction de la variable aléatoire.

Si $Y = \varphi(X)$ et si X possède une densité f alors que la densité de Y n'est pas calculable, les considérations précédentes sur l'espérance mathématique nous conduisent, afin d'évaluer les moments de Y, à calculer des expressions du type :

$$\mathrm{E}\{Y^n\} = \int_{-\infty}^{\infty} \varphi(x)^n f(x)\,dx\,, \tag{5.54}$$

sous réserve que de telles expressions existent.

5.6.3 Combinaison linéaire de variables aléatoires.

Soient $X_1, \ldots, X_n$ n variables aléatoires quelconques, et Y une variable aléatoire telle que :

$$Y = \sum_{i=1}^{n} a_i X_i \ . \tag{5.55}$$

Moyenne : l'espérance mathématique étant une fonctionnelle linéaire, on a :

$$\mathrm{E}\Big\{\sum_{i=1}^{n} a_i X_i\Big\} = \sum_{i=1}^{n} a_i \,\mathrm{E}\,\{X_i\} \tag{5.56}$$

Variance : on montre également que :

$$\mathrm{Var}(\sum_{i=1}^{n} a_i X_i) = \sum_{i=1}^{n} a_i^2 \,\mathrm{Var}(X_i) + 2\sum_{i=1}^{n}\sum_{j=i+1}^{n} a_i a_j \,\mathrm{Cov}(X_i, X_j) \tag{5.57}$$

et si les X_i sont deux à deux non corrélés on a pour $i \neq j \quad \mathrm{Cov}(X_i, X_j) = 0$ et :

$$\mathrm{Var}(\sum_{i=1}^{n} a_i X_i) = \sum_{i=1}^{n} a_i^2 \,\mathrm{Var}(X_i) \ . \tag{5.58}$$

Dans le cas particulier où les $a_i = 1$, cette dernière équation est connue sous le nom « d'égalité de Bienaymé ».

▶**Exemple 5.4.** *Gain d'un détecteur parfait.* Un détecteur de photons donne en sortie un nombre x exprimé en unités arbitraires, mais proportionnel au nombre n de photons détectés. La valeur x est appelée le nombre de « pas-codeurs ». Il faut plusieurs impacts de photons pour que le détecteur enregistre un pas-codeur et l'on demande le nombre de photons détectés par pas-codeur. On suppose que le détecteur n'introduit aucune source de bruit supplémentaire venant s'ajouter au bruit de photons.

Pour résoudre ce problème, on enregistre un échantillon de bruit à partir d'une source stationnaire. Soient $\bar{x}$ et Δx la moyenne et l'écart type de cet échantillon. Si n est le nombre de photons et α le nombre de photons par pas-codeur $n = \alpha x$ (le gain g est l'inverse de α) on a : $\bar{x} \equiv \mathrm{E}\,\{x\} = \mathrm{E}\,\{\frac{n}{\alpha}\} = \frac{1}{\alpha}\mathrm{E}\,\{n\}$ et $(\Delta x)^2 \equiv \mathrm{Var}(x) = \mathrm{Var}\,(\frac{n}{\alpha}) = \frac{1}{\alpha^2}\mathrm{Var}(n)$. L'émission de photons suivant une loi de Poisson, pour laquelle on a $\mathrm{Var}(n) = \mathrm{E}\,\{n\}$, il vient alors $\bar{x} = \frac{1}{\alpha}\mathrm{E}\,\{n\}$, $(\Delta x)^2 = \frac{1}{\alpha^2}\mathrm{E}\,\{n\}$. En éliminant $\mathrm{E}\,\{n\}$ on trouve :

$$\alpha = \frac{\bar{x}}{(\Delta x)^2} \tag{5.59}$$

Le facteur de conversion de pas-codeur en photons est donc égal au rapport de la moyenne sur la variance d'un échantillon de bruit, mesuré en pas-codeurs, issu d'une source stationnaire.

5.6.4 Moyenne et variance de la moyenne arithmétique.

Soient $(X_1, \ldots, X_n)$, n variables aléatoires suivant la même loi, de moyenne $\mathrm{E}\,\{X_i\} = \mu$, de variance $\mathrm{Var}(X_i) = \sigma^2$ et de coefficients de corrélation ρ_{ij}. La moyenne arithmétique M est une variable aléatoire définie par :

$$M = \frac{1}{n}\sum_{i=1}^{n} X_i \ . \tag{5.60}$$

C'est une expression identique à (5.55) à la condition de poser $a_i = \frac{1}{n}$. L'équation (5.56) nous permet alors d'obtenir l'espérance de M :

$$\mathrm{E}\{M\} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}\{X_i\} \quad \text{soit} \quad \mathrm{E}\{M\} = \mu\,. \tag{5.61}$$

Ce qui démontre que l'espérance mathématique de la moyenne arithmétique d'un échantillon est égale à la moyenne de la loi. Calculons maintenant la variance de la moyenne arithmétique M à l'aide de (5.57) :

$$\mathrm{Var}(M) = \sum_{i=1}^{n} \frac{1}{n^2}\mathrm{Var}(X_i) + \frac{2}{n^2}\sum_{i=1}^{n}\sum_{j=i+1}^{n} \mathrm{Cov}(X_i, X_j)\,,$$

soit, en posant $\mathrm{Cov}(X_i X_j) = \rho_{ij}\sigma^2$:

$$\mathrm{Var}(M) = \frac{\sigma^2}{n} + \frac{2\sigma^2}{n^2}\sum_{i=1}^{n}\sum_{j=i+1}^{n} \rho_{ij} \tag{5.62}$$

Si les variables aléatoires sont mutuellement non-corrélées ($i \neq j, \rho_{ij} = 0$) on a :

$$\mathrm{Var}(M) = \frac{\sigma^2}{n}\cdot \tag{5.63}$$

En particulier, si les X_i suivent une loi normale $\mathcal{N}(\mu, \sigma^2)$, la moyenne arithmétique M suit également une loi normale $\mathcal{N}(\mu, \sigma^2/n)$, de moyenne μ et de variance σ^2/n.

5.6.5 Changement de variables aléatoires linéaire nD.

Considérons maintenant le changement de variable linéaire $\boldsymbol{Y} = \boldsymbol{B}\boldsymbol{X}$, où $\boldsymbol{Y}$ et $\boldsymbol{X}$ représentent des vecteurs colonnes d'éléments respectivement $(Y_1, \ldots, Y_n)$ et $(X_1, \ldots, X_n)$ et $\boldsymbol{B}$ une matrice carrée (n, n). Soient $\boldsymbol{\mu_X}$ la moyenne de $\boldsymbol{X}$ et $\boldsymbol{V_X}$ sa matrice des variances-covariances. En utilisant la linéarité de l'espérance mathématique le calcul de la moyenne $\boldsymbol{\mu_Y}$ de $\boldsymbol{Y}$ est immédiat. On a :

$$\begin{aligned}\boldsymbol{\mu_Y} &= \mathrm{E}\{\boldsymbol{Y}\} = \mathrm{E}\{\boldsymbol{B}\boldsymbol{X}\} = \boldsymbol{B}\,\mathrm{E}\{\boldsymbol{X}\} \\ &= \boldsymbol{B}\boldsymbol{\mu_X}\ .\end{aligned}$$

Pour la matrice des variances-covariances $\boldsymbol{V_Y}$, on a :

$$\begin{aligned}\boldsymbol{V_Y} &= \mathrm{E}\{(\boldsymbol{Y} - \mathrm{E}\{\boldsymbol{Y}\})(\boldsymbol{Y} - \mathrm{E}\{\boldsymbol{Y}\})^t\} = \mathrm{E}\{\boldsymbol{Y}\boldsymbol{Y}^t\} - \mathrm{E}\{\boldsymbol{Y}\}\,\mathrm{E}\{\boldsymbol{Y}\}^t \\ &= \mathrm{E}\{\boldsymbol{B}\boldsymbol{X}\boldsymbol{X}^t\boldsymbol{B}^t\} - \mathrm{E}\{\boldsymbol{B}\boldsymbol{X}\}\,\mathrm{E}\{\boldsymbol{B}\boldsymbol{X}\}^t = \boldsymbol{B}\,\mathrm{E}\{\boldsymbol{X}\boldsymbol{X}^t\}\boldsymbol{B}^t - \boldsymbol{B}\,\mathrm{E}\{\boldsymbol{X}\}\,\mathrm{E}\{\boldsymbol{X}^t\}\boldsymbol{B}^t \\ &= \boldsymbol{B}(\mathrm{E}\{\boldsymbol{X}\boldsymbol{X}^t\} - \mathrm{E}\{\boldsymbol{X}\}\,\mathrm{E}\{\boldsymbol{X}^t\})\boldsymbol{B}^t \\ &= \boldsymbol{B}\boldsymbol{V_X}\boldsymbol{B}^t\ .\end{aligned}$$

Le tableau 5.2 résume ces résultats.

	Moyenne	Variance
X	μ	σ^2
aX	$a\mu$	$a^2\sigma^2$
$a_1X_1 + a_2X_2$	$a_1\mu_1 + a_2\mu_2$	$a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + 2a_1a_2\rho_{ij}\sigma_1\sigma_2$
$\boldsymbol{X}$	$\boldsymbol{\mu}$	$\boldsymbol{V}$
$\boldsymbol{BX}$	$\boldsymbol{B\mu}$	$\boldsymbol{BVB}^t$

TAB. 5.2: *Moyenne et variance ou matrice des variances-covariances des changements de variables aléatoires linéaires. La dernière ligne admet naturellement toutes les autres comme cas particuliers.*

5.6.6 Changement quasi-linéaire de variables aléatoires.

Soient les variables aléatoires $(X_1, \ldots, X_n)$, de moyennes μ_i et de variances σ_i^2. On définit une nouvelle variable aléatoire Y par l'intermédiaire du changement de variable $Y = \varphi(X_1, \ldots, X_n)$. On dira que ce changement de variable est quasi-linéaire, si la fonction φ est bien représentée par son approximation affine dans le domaine couvert en pratique par les variations des X_i. Dans ces conditions, on peut approximer la fonction φ par son développement de Taylor autour d'un point caractéristique de la répartition des X_i, par exemple, autour de la moyenne. On a alors :

$$\varphi(X_1, \ldots, X_n) \approx \varphi(\mu_1, \ldots, \mu_n) + \sum_{i=1}^{n} \left(\frac{\partial\varphi}{\partial x_i}\right)_{x_i=\mu_i} (X_i - \mu_i) \cdot \tag{5.64}$$

A l'aide de cette approximation on peut calculer la moyenne de Y, à condition naturellement que cette moyenne existe. En prenant la valeur moyenne de part et d'autre de l'expression (5.64), et en remarquant que $\mathrm{E}\{X_i - \mu_i\} = 0$, on obtient :

$$\mathrm{E}\{Y\} \approx \varphi(\mu_1, \ldots, \mu_n) \tag{5.65}$$

Dans la mesure où la variance de Y existe, on a également :

$$\mathrm{Var}(\varphi(X_1, \ldots, X_n)) \approx \sum_{i,j=1}^{n} \left(\frac{\partial\varphi}{\partial x_i}\right)_{x_i=\mu_i} \left(\frac{\partial\varphi}{\partial x_j}\right)_{x_j=\mu_j} \mathrm{Cov}(X_i, X_j)\,, \tag{5.66}$$

et si les variables aléatoires X_i sont non-corrélées on obtient la formule approchée :

$$\mathrm{Var}(\varphi(X_1, \ldots, X_n)) \approx \sum_{i=1}^{n} \left(\frac{\partial\varphi}{\partial x_i}\right)^2_{x_i=\mu_i} \mathrm{Var}(X_i)\,. \tag{5.67}$$

Cette formule est souvent appelée la « *formule de propagation des erreurs* ».

►**Exemple 5.5.** *Moyenne et variance approchées du module de deux variables aléatoires.* A partir du couple de variables aléatoires (X_1, X_2) on calcule la nouvelle variable aléatoire Y égale au module du couple, $Y = \sqrt{X_1^2 + X_2^2}$. Soient μ_1 et μ_2 les moyennes de X_1 et X_2. Supposons que les variations de X_1 et X_2 soient suffisamment faibles autour de leur moyenne pour nous permettre de représenter la fonction module par son approximation affine. On trouve en appliquant (5.65) :

$$\mathrm{E}\{Y\} \approx \sqrt{\mu_1^2 + \mu_2^2}\,. \tag{5.68}$$

Pour le calcul des variances, nous supposerons que les variables X_1 et X_2 sont non-corrélées et de variances σ_1^2 et σ_2^2. Les dérivées partielles de φ valent $\partial\varphi/\partial x_i = x_i(x_1^2 + x_2^2)^{-\frac{1}{2}}$, et par application de la formule (5.67), on trouve la variance de Y :

$$\mathrm{Var}\,(Y) \approx \frac{\mu_1^2\sigma_1^2 + \mu_2^2\sigma_2^2}{\mu_1^2 + \mu_2^2}\,. \tag{5.69}$$

5.7 Exercices et problèmes.

Exercice 5.1. Montrer que si X est une variable aléatoire à valeurs entières positives ($x \in \mathbb{N}^+$), alors on a :

$$\mathrm{E}\{X\} = \sum_{n=1}^{\infty} \Pr\{X \geq n\}\,.$$

Exercice 5.2. Montrer que si l'espérance mathématique $\mathrm{E}\{X\}$ d'une loi de fonction de répartition $F(x)$ existe alors elle est telle que :

$$\mathrm{E}\{X\} = \int_0^{\infty} (1 - F(x))\,dx - \int_{-\infty}^{0} F(x)\,dx\,. \tag{5.70}$$

Montrer qu'alors on peut redéfinir l'espérance mathématique comme la valeur μ telle que :

$$\int_{\mu}^{\infty} (1 - F(x))\,dx = \int_{-\infty}^{\mu} F(x)\,dx\,. \tag{5.71}$$

Exercice 5.3. L'écart absolu moyen de la variable aléatoire X, calculé autour de a, est défini par $\mathrm{E}\{|X-a|\}$, si cette valeur existe. Montrer que l'on a toujours $\mathrm{E}\{|X-a|\} \geq \mathrm{E}\{|X - x_{0.5}|\}$, où $x_{0.5}$ désigne la médiane de la loi suivie par X. Montrer qu'il n'y a égalité que lorsque $a = x_{0.5}$.

Problème 5.4. *Le problème des partis.* Deux joueurs sont convenus de jouer une partie en n points, c'est-à-dire que le premier qui marque n points remporte une certaine somme d'argent appelée *l'enjeu* de la partie. La partie comporte plusieurs tours, et à chaque tour, il n'y a qu'un et un seul joueur qui marque un point. Le résultat d'un tour est soumis au hasard et le jeu est équitable.

Supposons que, pour une raison fortuite, les deux joueurs sont obligés de se séparer alors que la partie n'est pas terminée. Le joueur P a obtenu $n-p$ points et le joueur Q en a obtenu $n-q$. Comment partager l'enjeu de façon équitable entre les deux joueurs, compte tenu des points déjà acquis ?

Ce problème de la partition équitable de l'enjeu est connu sous le nom de « problème des partis » et avait été proposé à la réflexion de Pascal par le Chevalier de Méré (voir Pascal : le triangle arithmétique [56]).

- Préciser ce que l'on doit entendre par « partage équitable » de l'enjeu.
- Si $P(p,q)$ désigne la proportion de l'enjeu qui revient au joueur P, montrer que cette valeur est définie par une suite récursive que l'on précisera.
- Démontrer la formule trouvée par Pascal lui-même qui est :

$$P(p,q) = \frac{1}{2^{p+q-1}} \sum_{k=0}^{q-1} C_{p+q-1}^{k}\,.$$

- Démontrer finalement que cette formule est aussi égale à :

$$P(p,q) = \frac{1}{B(p,q)} \int_0^{\frac{1}{2}} x^{p-1}(1-x)^{q-1}\,dx\,,$$

Chapitre 6

Lois normales

La loi normale ou loi de Gauss a été introduite par de Moivre en 1738, elle n'a été popularisé par Gauss qu'en 1809.

6.1 Loi normale à une dimension.

Une variable aléatoire X admet une loi normale, si elle possède une densité de probabilité $f(x)$ donnée par l'expression :

$$\boxed{f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp -\frac{(x-\mu)^2}{2\sigma^2}.} \tag{6.1}$$

C'est une loi à deux paramètres réels : μ et σ, μ est un paramètre de position et σ un paramètre d'échelle. Le graphe de la loi normale pour $\mu = 0$ et $\sigma = 1$ est donné par la figure 6.1. On note $\mathcal{N}(\mu, \sigma^2)$ une variable aléatoire qui suit la loi normale de paramètres μ et σ^2.

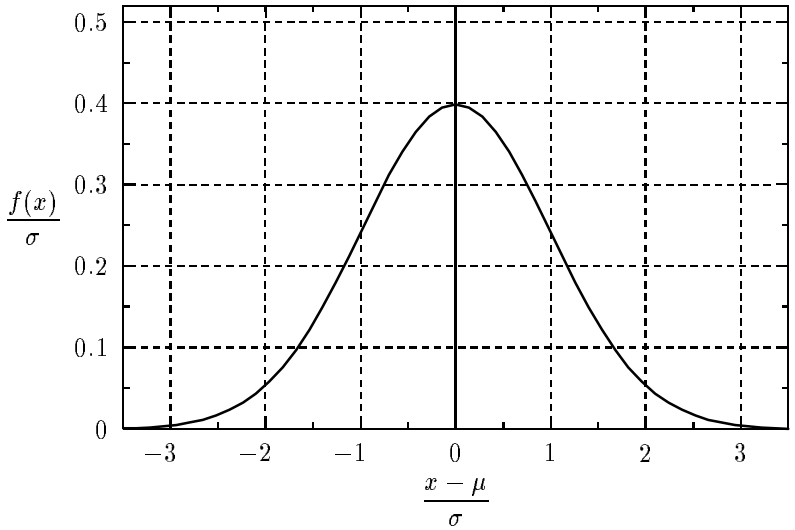


FIG. 6.1: *Densité de probabilité de la loi normale réduite.*

6.1.1 Fonction de répartition.

La fonction de répartition d'une variable aléatoire normale est donnée par l'expression :

$$F(x) = \Phi\Big(\frac{x-\mu}{\sigma}\Big) \quad \text{où} \quad \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}t^2}\, dt\,. \tag{6.2}$$

6.1.2 Fonction caractéristique.

La fonction caractéristique $Z(\omega)$ d'une variable aléatoire quelconque est par définition égale à $\mathrm{E}\{e^{i\omega X}\}$. Il vient pour une variable aléatoire normale :

$$\begin{aligned}
\mathrm{E}\{e^{i\omega X}\} &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{+\infty} e^{i\omega x} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}\, dx\,, \\
&= e^{i\mu\omega} \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{+\infty} e^{i\omega t} e^{-\frac{1}{2}(\frac{t}{\sigma})^2}\, dt\,, \\
&= e^{i\mu\omega} e^{-\frac{1}{2}\sigma^2\omega^2}\,,
\end{aligned}$$

soit :

$$Z(\omega) = \exp\{i\mu\omega - \tfrac{1}{2}\sigma^2\omega^2\}\,. \tag{6.3}$$

6.1.3 Caractéristiques numériques de la loi normale.

Moyenne et variance. Les paramètres μ et σ^2 sont respectivement la moyenne et la variance de la loi :

$$\mathrm{E}\{X\} = \mu, \quad \mathrm{Var}(X) = \sigma^2\,. \tag{6.4}$$

Le paramètre $\sigma > 0$ est l'écart type de la loi normale.

Moments centrés. Les moments centrés pour $r \geq 2$ sont donnés par les expressions suivantes :

$$\mu_{2r-1} = 0, \quad \mu_{2r} = \frac{(2r)!}{2^r r!}\sigma^{2r} = 1 \times 3 \times \cdots \times (2r-1)\sigma^{2r}\,. \tag{6.5}$$

Par exemple on a $\mu_3 = 0$ et $\mu_4 = 3\sigma^4$.

Asymétrie et aplatissement. De l'expression des moments centrés, on tire les coefficients d'asymétrie γ_1 et d'aplatissement γ_2 :

$$\gamma_1 = 0, \quad \gamma_2 = 0\,. \tag{6.6}$$

Le coefficient d'aplatissement, tel qu'il apparaît dans l'équation (2.33), a été défini de façon à être nul pour la loi normale.

Autres caractéristiques numériques. La largeur à mi-hauteur de la loi normale est donnée par l'expression :

$$\text{FWHM} = 2\sigma\sqrt{2\ln 2} \approx 2.3548\sigma\,. \tag{6.7}$$

L'écart absolu moyen e est égal à :

$$e = \int_{-\infty}^{\infty} |x-\mu|\, dF = \sqrt{\frac{2}{\pi}}\sigma \approx 0.79788\sigma\,. \tag{6.8}$$

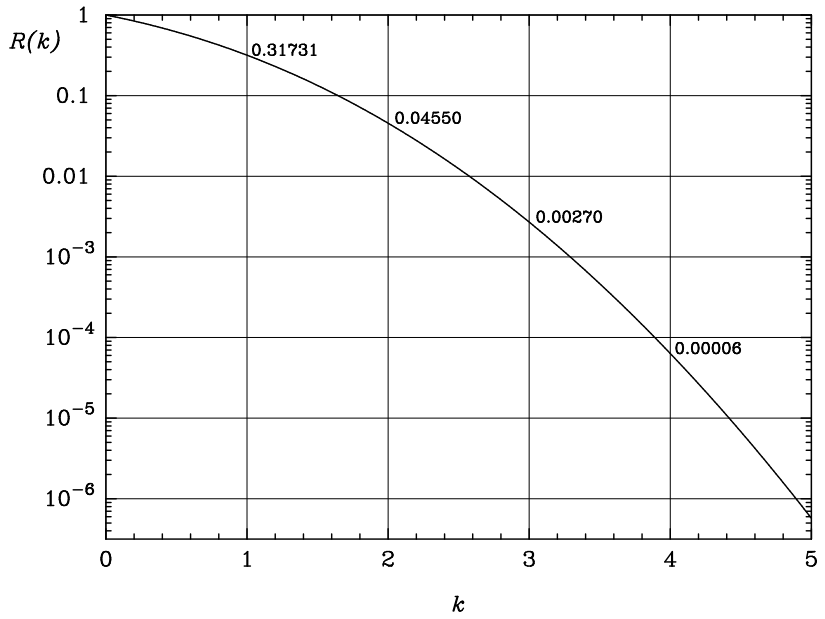


FIG. 6.2: *Fonction d'erreur résiduelle $R(k)$ de la loi normale. Cette fonction donne la probabilité pour qu'une variable aléatoire normale s'écarte de sa moyenne de plus que k fois son écart type. Par définition on a $R(k) = 1-[\Phi(k)-\Phi(-k)]$, où Φ est la fonction de répartition de la loi normale réduite.*

6.1.4 Quelques propriétés de la loi normale.

Loi normale réduite. La variable aléatoire $Y = (X-\mu)/\sigma$ est une variable aléatoire normale de moyenne nulle et de variance unité, elle est appelée variable aléatoire normale réduite et suit une loi normale réduite $\mathcal{N}(0,1)$.

Quantiles et intervalle de confiance de la loi normale. Pour un α donné, cherchons à résoudre les équations du type :

$$\Pr\Big\{\frac{|X-\mu|}{\sigma} \leq Q_{\frac{\alpha}{2}}\Big\} = 1-\alpha$$

La quantité $Q_{\frac{\alpha}{2}}$ est un quantile d'ordre $\alpha/2$ de la loi normale réduite. L'intervalle $[-Q_{\frac{\alpha}{2}}, Q_{\frac{\alpha}{2}}]$ est appelé l'intervalle inter-quantile d'ordre α. On trouve $Q_{\frac{\alpha}{2}}$ à l'aide de la fonction de répartition Φ de la loi normale réduite :

$$Q_{\frac{\alpha}{2}} = \Phi^{-1}\Big(1-\frac{\alpha}{2}\Big)\,. \tag{6.9}$$

Parmi les intervalles inter-quantiles on distingue l'intervalle $[-Q_{0.25}, Q_{0.25}]$ appelé « *intervalle inter-quartile* ». On préfère parfois poser $\gamma = 1-\alpha$ et parler de *l'intervalle de confiance au niveau* γ : $[-Q_{\frac{1}{2}-\frac{\gamma}{2}}, Q_{\frac{1}{2}-\frac{\gamma}{2}}]$. La table 6.1 donne les valeurs de $Q_{\frac{1}{2}-\frac{\gamma}{2}}$ pour des valeurs typiques de γ.

100γ	50.0%	68.3%	90.0%	95.0%
$Q_{\frac{1}{2}-\frac{\gamma}{2}}$	0.6745	1	1.6449	1.9600
100γ	95.4%	99.0%	99.7%	99.9%
$Q_{\frac{1}{2}-\frac{\gamma}{2}}$	2	2.5758	3	3.2905

TAB. 6.1: *Quantiles permettant de calculer un intervalle de confiance au niveau γ de la loi normale réduite. Par exemple, au niveau de confiance $\gamma = 99.7\%$ est associé l'intervalle $[-3, 3]$. En d'autres termes, une valeur x issue d'une loi normale réduite sera dans 99.7% des cas comprise entre -3 et 3.*

Somme de variables aléatoires normales. La somme de deux variables aléatoires normales indépendantes X_1 et X_2 de moyennes μ_1, μ_2 et de variances σ_1^2, σ_2^2 est également une variable aléatoire normale, de moyenne $\mu_1 + \mu_2$ et de variance $\sigma_1^2 + \sigma_2^2$. La réciproque est également vraie : si la somme de deux variables aléatoires indépendantes suit une loi normale alors ces deux variables suivent aussi un loi normale (voir H. Cramér 1936 [17]). On a :

$$X_1 \in \mathcal{N}(\mu_1, \sigma_1^2), X_2 \in \mathcal{N}(\mu_2, \sigma_2^2) \iff X_1 + X_2 \in \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)\,. \tag{6.10}$$

La loi normale est dite « indéfiniment divisible », ce résultat s'étend à un nombre quelconque de variables aléatoires normales.

6.2 Loi normale à 2 dimensions.

Un vecteur aléatoire $\boldsymbol{X} = (X_1, X_2)$ admet une loi normale 2D s'il possède une densité de probabilité donnée par l'expression :

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2(1-\rho^2)^{\frac{1}{2}}} \exp\left\{-\frac{1}{2(1-\rho^2)} \left[\frac{(x_1-\mu_1)^2}{\sigma_1^2} - \frac{2\rho(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2}\right]\right\} \tag{6.11}$$

C'est une loi à cinq paramètres : $\mu_1, \mu_2, \sigma_1, \sigma_2$ et ρ, ($\sigma_1, \sigma_2 > 0$ et $-1 \leq \rho \leq 1$). Nous ne considérerons que la loi non-dégénérée pour laquelle $\sigma_1\sigma_2(1-\rho^2) \neq 0$, ce qui revient à dire que le coefficient ρ n'est ni égal à 1, ni égal à -1 et que $\sigma_1, \sigma_2 \neq 0$.

6.2.1 Fonction caractéristique 2D.

Par définition on a $Z(\omega_1, \omega_2) = \mathrm{E}\{\exp i(\omega_1 X_1 + \omega_2 X_2)\}$, il vient :

$$Z(\omega_1, \omega_2) = \exp\{-\tfrac{1}{2}(\sigma_1^2\omega_1^2 + 2\rho\sigma_1\sigma_2 + \sigma_2^2\omega_2^2)\} \exp\{i(\omega_1\mu_1 + \omega_2\mu_2)\}\,. \tag{6.12}$$

6.2.2 Lois conditionnelles.

Nous n'envisagerons ici que les lois conditionnelles suivant une coupe parallèle à l'axe x_1 ou à l'axe x_2. On cherche, par exemple, la loi suivie par X_1 lorsque X_2 est connue et vaut x_2. D'après les résultats du paragraphe 3.1.9, la densité

de cette loi conditionnelle est égale à $f(x_1, x_2)$ envisagée comme fonction de la seule variable x_1 et normalisée par intégration sur x_1. Si $f_{X_1|X_2}$ désigne cette densité, on a $f_{X_1|X_2}(x_1|X_2 = x_2) \propto f(x_1, x_2)$. Il vient, en regroupant dans la constante d'intégration les termes qui ne dépendent pas de x_1 :

$$\begin{aligned} f_{X_1|X_2}(x_1|X_2 = x_2) &\propto \exp\Big\{-\frac{1}{2(1-\rho^2)}\Big[\frac{(x_1-\mu_1)^2}{\sigma_1^2} - 2\rho\frac{(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2}\Big]\Big\} \\ &\propto \exp\Big\{-\frac{1}{2(1-\rho^2)}\Big[\frac{(x_1-\mu_1)}{\sigma_1} - \rho\frac{(x_2-\mu_2)}{\sigma_2}\Big]^2\Big\} \\ &\propto \exp\Big\{-\frac{1}{2\sigma_1^2(1-\rho^2)}\big[(x_1-\mu_1) - \rho\frac{\sigma_1}{\sigma_2}(x_2-\mu_2)\big]^2\Big\} \\ &\propto \exp\Big\{-\frac{1}{2\sigma_1^2(1-\rho^2)}\big[x_1 - (\mu_1 + \rho\frac{\sigma_1}{\sigma_2}(x_2-\mu_2))\big]^2\Big\}, \end{aligned}$$

ce qui montre que la loi conditionnelle $f_{X_1|X_2}$ est normale, de moyenne $\mu_1 + \rho\sigma_1(x_2-\mu_2)/\sigma_2$ et de variance $\sigma_1^2(1-\rho^2)$. Pour une loi normale, la constante de normalisation est égale à l'inverse de son écart type multiplié par $\sqrt{2\pi}$. Il vient alors :

$$f_{X_1|X_2}(x_1|X_2 = x_2) = \frac{1}{\sqrt{2\pi}\sigma_1(1-\rho^2)^{\frac{1}{2}}}\exp\Big\{-\frac{[x_1 - (\mu_1 + \rho\frac{\sigma_1}{\sigma_2}(x_2-\mu_2))]^2}{2\sigma_1^2(1-\rho^2)}\Big\}. \tag{6.13}$$

L'expression de la loi $f_{X_2|X_1}$ de X_2 sachant que $X_1 = x_1$ s'obtient en permutant les indices 1 et 2 dans l'équation précédente.

6.2.3 Caractéristiques numériques de la loi normale 2D.

Moyenne. C'est un vecteur colonne :

$$\mathrm{E}\{\boldsymbol{X}\} = \begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}. \tag{6.14}$$

Moyennes conditionnelles. D'après les calculs précédents nous avons :

$$\mathrm{E}\{X_2|X_1 = x_1\} = \mu_2 + \frac{\rho\sigma_2}{\sigma_1}(x_1-\mu_1), \tag{6.15}$$

$$\mathrm{E}\{X_1|X_2 = x_2\} = \mu_1 + \frac{\rho\sigma_1}{\sigma_2}(x_2-\mu_2). \tag{6.16}$$

Les droites d'équations $x_1 = \mu_1 + \frac{\rho\sigma_1}{\sigma_2}(x_2-\mu_2)$ et $x_2 = \mu_2 + \frac{\rho\sigma_2}{\sigma_1}(x_1-\mu_1)$ sont les droites de régression, de X_2 par rapport à X_1, et de X_1 par rapport à X_2, (voir figure 6.4).

Matrice des variances-covariances. C'est une matrice $(2,2)$ dont les valeurs sont données par l'expression :

$$\boldsymbol{V} = \begin{pmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix}, \qquad \boldsymbol{V}^{-1} = \frac{1}{1-\rho^2}\begin{pmatrix}\frac{1}{\sigma_1^2} & \frac{-\rho}{\sigma_1\sigma_2}\\ \frac{-\rho}{\sigma_1\sigma_2} & \frac{1}{\sigma_2^2}\end{pmatrix}. \tag{6.17}$$

Variances conditionnelles. Ces variances sont indépendantes du niveau où l'on effectue la coupe :

$$\mathrm{Var}(X_1|X_2 = x_2) = \sigma_1^2(1-\rho^2), \quad \mathrm{Var}(X_2|X_1 = x_1) = \sigma_2^2(1-\rho^2)\,. \tag{6.18}$$

Coefficient de corrélation. Le coefficient de corrélation est égal au paramètre ρ de la loi :

$$\mathrm{corr}(X_1, X_2) = \rho\,. \tag{6.19}$$

6.2.4 Forme quadratique associée.

L'argument de l'exponentielle de l'équation (6.11) est une forme quadratique $Q(x_1, x_2)$:

$$Q(x_1, x_2) = \frac{1}{(1-\rho^2)}\left[\frac{(x_1-\mu_1)^2}{\sigma_1^2} - \frac{2\rho(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2}\right]. \tag{6.20}$$

Cette forme est homogène pour les variables $(x_1 - \mu_1)$ et $(x_2 - \mu_2)$. Définissons le vecteur $(\boldsymbol{x} - \boldsymbol{\mu})$ comme étant un vecteur colonne de composantes $(x_i - \mu_i)$. Soit $\boldsymbol{A}$ la matrice (symétrique) des demi-dérivées secondes de Q :

$$\boldsymbol{A} = \frac{1}{2}\begin{pmatrix} \dfrac{\partial^2 Q}{\partial x_1^2} & \dfrac{\partial^2 Q}{\partial x_1 \partial x_2} \\ \dfrac{\partial^2 Q}{\partial x_2 \partial x_1} & \dfrac{\partial^2 Q}{\partial x_2^2} \end{pmatrix}. \tag{6.21}$$

La théorie des formes quadratiques nous apprend que la forme Q peut être mise sous la forme matricielle :

$$Q(x_1, x_2) = (\boldsymbol{x} - \boldsymbol{\mu})^t \boldsymbol{A} (\boldsymbol{x} - \boldsymbol{\mu})\,. \tag{6.22}$$

Dans cette écriture le vecteur $(\boldsymbol{x} - \boldsymbol{\mu})^t$ est un vecteur ligne obtenu par transposition de $(\boldsymbol{x} - \boldsymbol{\mu})$. Il est possible de réduire cette forme quadratique à sa forme canonique, grâce à la transformation linéaire $\boldsymbol{U}$ qui diagonalise $\boldsymbol{A}$. La matrice $\boldsymbol{U}$ est formée des vecteurs propres de $\boldsymbol{A}$, écrits sous forme de colonnes et juxtaposés de façon à obtenir une matrice carrée. La matrice $\boldsymbol{A}$ étant symétrique, elle est effectivement diagonalisable. De plus ses vecteurs propres sont orthogonaux et il suffit de les normer afin que la matrice $\boldsymbol{U}$ soit orthonormée de façon à ce que l'on ait $\boldsymbol{U}^{-1} = \boldsymbol{U}^t$.

Il est aisé de constater que $\boldsymbol{A} = \boldsymbol{V}^{-1}$. La matrice $\boldsymbol{A}$ en tant qu'inverse de la matrice définie positive $\boldsymbol{V}$ est elle-même définie positive. Soient $\lambda_1^2 \geq \lambda_2^2$ les valeurs propres, nécessairement positives, de $\boldsymbol{V}$. Les valeurs propres de $\boldsymbol{A}$ sont alors égales à $1/\lambda_1^2$ et $1/\lambda_2^2$. Soit $\boldsymbol{\Lambda}^2$ la matrice (diagonale) des valeurs propres de $\boldsymbol{V}$. On a :

$$\boldsymbol{\Lambda}^{-2} = \boldsymbol{U}^t \boldsymbol{A} \boldsymbol{U} \quad \text{et} \quad \boldsymbol{U}\boldsymbol{\Lambda}^{-2}\boldsymbol{U}^t = \boldsymbol{A}. \tag{6.23}$$

La matrice $\boldsymbol{U}$ définit un changement de base. Soit $\boldsymbol{y}$ un vecteur colonne représentant les coordonnées d'un vecteur dans cette nouvelle base après translation d'un vecteur $\boldsymbol{\mu}$. On a par définition de la matrice de changement de

base $(\boldsymbol{x}-\boldsymbol{\mu}) = \boldsymbol{U}\boldsymbol{y}$. On obtient alors $(\boldsymbol{x}-\boldsymbol{\mu})^t \boldsymbol{A}(\boldsymbol{x}-\boldsymbol{\mu}) = \boldsymbol{y}^t \boldsymbol{\Lambda}^{-2}\boldsymbol{y}$. La forme quadratique (6.20) s'écrit dans cette nouvelle base :

$$Q(x_1, x_2) = \frac{y_1^2}{\lambda_1^2} + \frac{y_2^2}{\lambda_2^2}\,. \tag{6.24}$$

On calcule aisément les valeurs propres λ_1^2 et λ_2^2 de $\boldsymbol{V}$:

$$\lambda_{1,2}^2 = \tfrac{1}{2}(\sigma_1^2 + \sigma_2^2 \pm [(\sigma_1^2 - \sigma_2^2)^2 + 4\rho^2\sigma_1^2\sigma_2^2]^{\frac{1}{2}})\,. \tag{6.25}$$

Avec la convention $\lambda_1^2 \geq \lambda_2^2$, λ_1^2 correspond au signe $+$ et λ_2^2 au signe $-$. Soit $g(y_1, y_2)$ la densité de probabilité de la loi normale exprimée avec les nouvelles variables $\boldsymbol{y}$. La condition $\int g(\boldsymbol{y}) = 1$ permet de calculer la constante de normalisation et l'on obtient :

$$g(y_1, y_2) = \frac{1}{2\pi\lambda_1\lambda_2} \exp\left\{-\frac{1}{2}\left(\frac{y_1^2}{\lambda_1^2} + \frac{y_2^2}{\lambda_2^2}\right)\right\}\,. \tag{6.26}$$

Le changement de base correspond au changement de variables aléatoires $(\boldsymbol{X} - \boldsymbol{\mu}) = \boldsymbol{U}\boldsymbol{Y}$. Nous venons donc de montrer que $\boldsymbol{Y}$ possède la densité (6.26) et que ses composantes Y_1, Y_2 sont normales, indépendantes, de moyennes nulles et ont pour variances λ_1^2 et λ_2^2. Calculons maintenant la matrice $\boldsymbol{U}$ qui diagonalise $\boldsymbol{A}$. En tant que matrice unitaire dans $\mathbb{R}^2$, $\boldsymbol{U}$ est une matrice de rotation d'angle ϕ :

$$\boldsymbol{U} = \begin{pmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{pmatrix}\,. \tag{6.27}$$

Calculons les éléments de cette matrice avec la convention que $-\frac{\pi}{2} < \phi \leq \frac{\pi}{2}$, et que le vecteur propre $(\cos\phi, \sin\phi)$ correspond à la valeur propre λ_1^2. On trouve alors, à une constante multiplicative près :

$$\begin{pmatrix} \cos\phi \\ \sin\phi \end{pmatrix} \propto \begin{pmatrix} \lambda_1^2 - \sigma_2^2 \\ \rho\sigma_1\sigma_2 \end{pmatrix}\,. \tag{6.28}$$

D'un point de vue numérique, cette formule est valable dans la plupart des cas, sauf si $\lambda_1 \approx \sigma_2^2$ avec $\rho \approx 0$. Dans ce cas il vaut mieux employer la formule :

$$\begin{pmatrix} \cos\phi \\ \sin\phi \end{pmatrix} \propto \begin{pmatrix} \rho\sigma_1\sigma_2 \\ \lambda_1^2 - \sigma_1^2 \end{pmatrix}\,. \tag{6.29}$$

On peut également utiliser la formule :

$$\tan 2\phi = \frac{2\rho\sigma_1\sigma_2}{\sigma_1^2 - \sigma_2^2}, \tag{6.30}$$

qui donne deux valeurs de l'angle ϕ différant de $\frac{\pi}{2}$ et correspondant au grand axe et au petit axe de l'ellipse, mais cette formule, à elle seule, ne permet pas de les distinguer.

6.2.5 Ellipses d'égale probabilité.

La réduction de la forme quadratique $(\boldsymbol{x}-\boldsymbol{\mu})^t \boldsymbol{V}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$ à sa forme canonique (6.24) a bien mis en évidence que cette forme était définie positive. L'équation $Q(x_1, x_2) = k^2$ est donc l'équation d'une ellipse centrée sur la moyenne $\boldsymbol{\mu}$.

Le nouveau couple de coordonnées (y_1, y_2), défini par la matrice de changement de base $\boldsymbol{U}$, correspond à deux axes passant par le centre de l'ellipse et confondus avec le grand axe et le petit axe de l'ellipse. Le long de cette ellipse, la densité de probabilité de la loi normale 2D est constante et vaut :

$$f(\boldsymbol{x}) = \frac{1}{2\pi\sigma_1\sigma_2} \exp\{-\tfrac{1}{2}k^2\} \,. \tag{6.31}$$

Choisissons la constante $k^2 = 1$. L'ellipse ainsi obtenue est appelée *ellipse de corrélation*, voir figure 6.3.

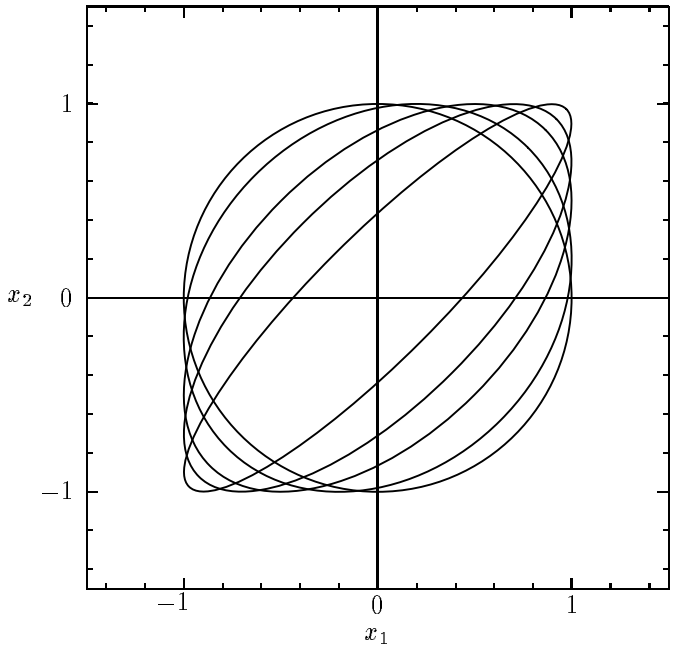


FIG. 6.3: *Ellipses de corrélation de la loi normale 2D pour* $\mu_1 = \mu_2 = 0$, $\sigma_1 = \sigma_2 = 1$ *et de coefficient de corrélation prenant successivement les valeurs* $\rho = 0.0, 0.2, 0.5, 0.7$ *et* 0.9. *Ces ellipses restent inscrites dans le rectangle (ici un carré) de dispersion. Les ellipses de corrélation contiennent la probabilité* $\gamma = 1 - \exp\{-k^2/2\}$, *[voir plus loin l'équation* (6.45)*]. Ici, avec* $k^2 = 1$, *on a* $\gamma = 39.3\%$.

Convenons d'appeler *rectangle de dispersion*, le rectangle de côtés parallèles aux axes et circonscrit à l'ellipse de corrélation. Nous allons montrer que la longueur des côtés de ce rectangle est égale à $2\sigma_1$ et $2\sigma_2$.

Pour cela, nous devons chercher le lieu des points stationnaires $dx_1 = 0$ et $dx_2 = 0$, sur l'ellipse $Q(x_1, x_2) = 1$. Pour simplifier on supposera que l'on a translaté les axes sur le centre de l'ellipse et qu'alors $\mu_1 = \mu_2 = 0$ de façon a ce que la forme Q soit homogène pour les variables x_1 et x_2. Sur l'ellipse, la forme Q est constante et donc :

$$dQ = \frac{\partial Q}{\partial x_1} dx_1 + \frac{\partial Q}{\partial x_2} dx_2 = 0 \,. \tag{6.32}$$

Cherchons, par exemple, les points tels que $dx_1 = 0$, dx_2 étant non-nul, la condition (6.32) impose $\partial Q/\partial x_2 = 0$. Remplaçons cela dans la forme quadratique qui, étant homogène, peut s'écrire :

$$Q = \frac{1}{2}\Big(x_1\frac{\partial Q}{\partial x_1} + x_2\frac{\partial Q}{\partial x_2}\Big) = 1\,. \tag{6.33}$$

Il vient $\frac{1}{2}x_1\partial Q/\partial x_1 = 1$. Nous devons à présent résoudre le système :

$$\begin{aligned}\frac{1}{2}x_1\frac{\partial Q}{\partial x_1} &= \frac{x_1}{1-\rho^2}\Big(\frac{x_1}{\sigma_1^2} - \rho\frac{x_2}{\sigma_1\sigma_2}\Big) = 1\,,\\ \frac{\partial Q}{\partial x_2} &= \frac{2}{1-\rho^2}\Big(-\rho\frac{x_1}{\sigma_1\sigma_2} + \frac{x_2}{\sigma_2^2}\Big) = 0\,.\end{aligned}$$

Ce système a pour solution :

$$x_1 = \pm\sigma_1\,, \tag{6.34}$$

$$x_2 = \rho\frac{\sigma_2}{\sigma_1}x_1 = \pm\rho\sigma_2\,. \tag{6.35}$$

De la même façon on trouverait pour le lieu des points tels que $dx_2 = 0$:

$$x_1 = \rho\frac{\sigma_1}{\sigma_2}x_2 = \pm\rho\sigma_1\,, \tag{6.36}$$

$$x_2 = \pm\sigma_2\,. \tag{6.37}$$

Notons que les équations (6.35) et (6.37) sont identiques aux droites de régression (6.15) et (6.16). La figure 6.4 présente une interprétation géométrique des propriétés qui viennent d'être démontrées.

Contenu en probabilité des ellipses d'égale probabilité.

Cherchons maintenant le contenu en probabilité P_{k^2} des ellipses $Q(x_1, x_2) = k^2$ que nous venons de définir. Il faut pour cela évaluer l'intégrale suivante :

$$P_{k^2} = \Pr\left\{X_1, X_2 | Q(X_1, X_2) \leq k^2\right\} = \iint\limits_{Q(u,v)\leq k^2} f(u,v)\,du dv\,. \tag{6.38}$$

La quantité P_{k^2} étant la probabilité d'un certain événement A, elle ne dépend pas d'un choix particulier des variables aléatoires choisies pour le représenter à la condition qu'il y ait bijection entre ces variables aléatoires. Cette dernière condition est automatiquement remplie lorsque le déterminant de la matrice de changement de base n'est pas nul.

Effectuons maintenant une translation de vecteur $\boldsymbol{\mu}$ suivie d'un changement de base linéaire de matrice $\boldsymbol{U\Lambda}$. Les nouvelles coordonnées $\boldsymbol{y}$ sont, par définition, telles que $(\boldsymbol{x} - \boldsymbol{\mu}) = \boldsymbol{U\Lambda y}$. La forme quadratique devient alors :

$$(\boldsymbol{x}-\boldsymbol{\mu})^t\boldsymbol{V}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) = (\boldsymbol{U\Lambda y})^t\boldsymbol{V}^{-1}(\boldsymbol{U\Lambda y}) \tag{6.39}$$

$$= \boldsymbol{y}^t\boldsymbol{\Lambda}\underbrace{\boldsymbol{U}^t\boldsymbol{V}^{-1}\boldsymbol{U}}_{\boldsymbol{\Lambda}^{-2}}\boldsymbol{\Lambda y} \tag{6.40}$$

$$= \boldsymbol{y}^t\boldsymbol{y} = y_1^2 + y_2^2 = k^2 \tag{6.41}$$

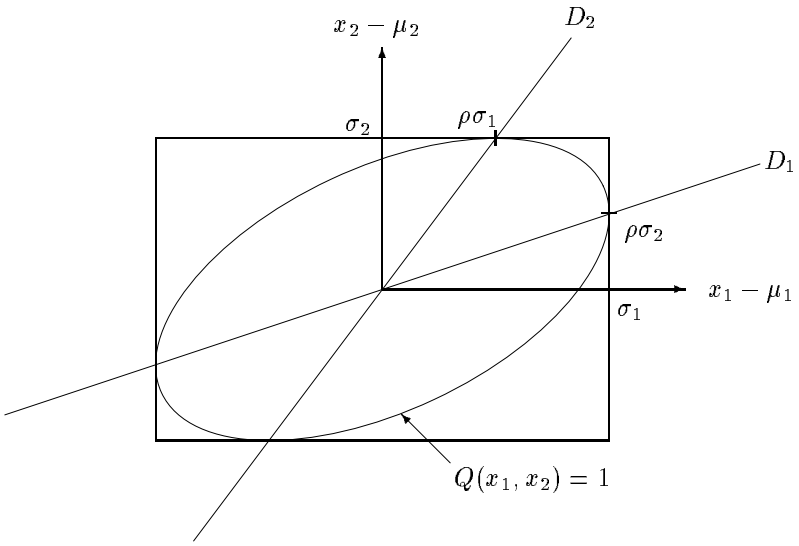


FIG. 6.4: *Interprétation géométrique du rectangle de dispersion associé à une loi normale 2D. Ce rectangle est circonscrit à l'ellipse de corrélation* $Q(x_1, x_2) = 1$ *et a pour côtés* $2\sigma_1$ *et* $2\sigma_2$*. La droite* D_1 *est la droite de régression de* ξ_2 *par rapport à* ξ_1 *et* D_2 *la droite de régression de* ξ_1 *par rapport à* ξ_2*. Sur ce graphique on a* $\sigma_1 = 3$, $\sigma_2 = 2$ *et* $\rho = 0.5$.

Avec ces nouvelles coordonnées, la densité de probabilité vaut maintenant :

$$f_y(y_1, y_2) = \frac{1}{2\pi} \exp\{-\tfrac{1}{2}(y_1^2 + y_2^2)\} \,. \tag{6.42}$$

Il est à présent facile d'évaluer P_{k^2} :

$$P_{k^2} = \frac{1}{2\pi} \iint\limits_{u^2+v^2 \le k^2} \exp\{-\tfrac{1}{2}(u^2 + v^2)\} \, du dv \,. \tag{6.43}$$

Le changement de variable $u = r\cos\varphi, v = r\sin\varphi$, nous permet finalement d'écrire :

$$P_{k^2} = \frac{1}{2\pi} \int_0^{2\pi} d\varphi \int_0^k \exp\{-\tfrac{1}{2}r^2\} r \, dr \,, \tag{6.44}$$

d'où la probabilité cherchée :

$$P_{k^2} = 1 - e^{-\frac{1}{2}k^2} \,. \tag{6.45}$$

Cette quantité ne dépend pas du coefficient de corrélation ρ ni d'aucun autre paramètre de la loi normale. Un tel comportement était prévisible, car P_{k^2}, en tant que probabilité, doit être invariante par translation et changement d'échelle (qui sont des bijections) ce qui la rend indépendante de μ_1, μ_2 et de σ_1, σ_2, mais aussi par rotation (qui est aussi une bijection) ce qui la rend indépendante de ρ. Par ailleurs, sous ces transformations linéaires, une forme quadratique reste une forme quadratique, et on aurait pu se placer dans le cas $\mu_1 = \mu_2 = 0$, $\sigma_1 = \sigma_2 = 1$, $\rho = 0$, ce qui menait directement à l'équation (6.43).

6.2.6 Forme matricielle de la loi normale 2D.

Nous avons vu que l'on pouvait mettre l'argument de l'exponentielle de la loi normale sous la forme :

$$-\tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^t\boldsymbol{V}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\,, \tag{6.46}$$

où $\boldsymbol{V}$ est la matrice des variances-covariances du couple de variables aléatoires X_1, X_2. Nous avons également vu que le changement de variable $(\boldsymbol{x}-\boldsymbol{\mu}) = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{y}$, où $\boldsymbol{U}$ est la matrice des vecteurs propres de $\boldsymbol{V}$ et $\boldsymbol{\Lambda}$ la matrice diagonale formée des racines carrées des valeurs propres de $\boldsymbol{V}$, transformait cet argument en une simple somme de carrés. Avec ces nouvelles variables $\boldsymbol{y}$, la loi normale s'écrit maintenant :

$$f_{\boldsymbol{Y}}(y_1, y_2) = \frac{1}{2\pi}e^{-\frac{1}{2}\boldsymbol{y}^t\boldsymbol{y}}\,. \tag{6.47}$$

Calculons ce que devient la constante de normalisation au cours du changement de base inverse. La probabilité doit être conservée, de telle façon que :

$$f_{\boldsymbol{Y}}(y_1, y_2)dy_1dy_2 = f_{\boldsymbol{X}}(x_1, x_2)dx_1dx_2\,, \tag{6.48}$$

mais $dy_1dy_2 = |J|dx_1dx_2$, où J est le jacobien du changement de base. Le changement de base est ici linéaire et le jacobien est alors égal au déterminant de la matrice de changement de base :

$$J^{-1} = \det(\boldsymbol{U}\boldsymbol{\Lambda}) = \det\boldsymbol{\Lambda}\,, \tag{6.49}$$

$$\det\boldsymbol{\Lambda} = \sqrt{\det(\boldsymbol{U}^t\boldsymbol{V}\boldsymbol{U})} = \sqrt{\det\boldsymbol{V}}\,, \tag{6.50}$$

d'où la forme matricielle de la loi normale 2D :

$$f(x_1, x_2) = \frac{1}{2\pi(\det\boldsymbol{V})^{\frac{1}{2}}}\exp\{-\tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^t\boldsymbol{V}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\}\,. \tag{6.51}$$

6.2.7 Lois marginales.

Pour obtenir les densités de probabilité des lois marginales, il faut intégrer $f(u, v)$ sur les demi-plans $\mathcal{D}_{x_1}$ et $\mathcal{D}_{x_2}$ définis respectivement par les équations $u \leq x_1$ et $v \leq x_2$ puis dériver par rapport aux variables x_1 et x_2 :

$$f_{X_1}(x_1) = \frac{d}{dx_1}\iint_{\mathcal{D}_{x_1}} f(u, v)\,dudv\,, \tag{6.52}$$

$$f_{X_2}(x_2) = \frac{d}{dx_2}\iint_{\mathcal{D}_{x_2}} f(u, v)\,dudv\,. \tag{6.53}$$

Il n'est cependant pas nécessaire de calculer ces intégrales, on peut trouver les densités marginales en utilisant les relations (3.40) qui lient entre elles les densités 2D, conditionnelles et marginales. Pour f_{X_1} par exemple on a :

$$f_{X_1}(x_1) = \frac{f(x_1, x_2)}{f_{X_2|X_1}(x_2|X_1 = x_1)}\,. \tag{6.54}$$

La loi conditionnelle $f_{X_2|X_1}$ est donnée par une expression analogue à l'équation (6.13) et il vient en réarrangeant un peu les termes de cette expression :

$$f_{X_1}(x_1) = \frac{\sqrt{2\pi}\sigma_2(1-\rho^2)^{\frac{1}{2}}}{2\pi\sigma_1\sigma_2(1-\rho^2)^{\frac{1}{2}}} \times$$
$$\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(x_1-\mu_1)^2}{\sigma_1^2} - \frac{2\rho(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2}\right]\right\} \times$$
$$\exp\left\{\frac{1}{2(1-\rho^2)}\left[\frac{(x_2-\mu_2)}{\sigma_2} - \rho\frac{(x_1-\mu_1)}{\sigma_1}\right]^2\right\}$$

La plupart des termes des exponentielles disparaissent deux à deux et il reste :

$$f_{X_1}(x_1) = \frac{1}{\sqrt{2\pi}\sigma_1}\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - \rho^2\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2\right]\right\}.$$

D'où l'expression de la loi marginale de X_1 :

$$f_{X_1}(x_1) = \frac{1}{\sqrt{2\pi}\sigma_1}\exp\left\{-\frac{1}{2}\frac{(x_1-\mu_1)^2}{\sigma_1^2}\right\}. \tag{6.55}$$

De la même façon on trouverait pour la loi marginale de X_2 :

$$f_{X_2}(x_2) = \frac{1}{\sqrt{2\pi}\sigma_2}\exp\left\{-\frac{1}{2}\frac{(x_2-\mu_2)^2}{\sigma_2^2}\right\}, \tag{6.56}$$

ce qui démontre que les lois marginales de la loi normale 2D sont des lois normales 1D de même moyenne et de même variance que la loi 2D. Notons que ces lois ne dépendent pas du coefficient de corrélation ρ.

6.3 Loi normale à n dimensions.

Un vecteur aléatoire $\boldsymbol{X} = (X_1, \ldots, X_n)$ suit une loi normale à n dimensions (nD) s'il possède une densité de probabilité donnée par l'expression :

$$\boxed{f(\boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{n}{2}}(\det \boldsymbol{V})^{\frac{1}{2}}}\exp\{-\tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^t\boldsymbol{V}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\},} \tag{6.57}$$

où $\boldsymbol{\mu}$ désigne un vecteur colonne à n composantes : $(\mu_1, \ldots, \mu_n)$ et $\boldsymbol{V}$ une matrice carrée symétrique définie positive possédant n lignes et n colonnes. C'est une loi à $n(n+3)/2$ paramètres : n paramètres pour $\boldsymbol{\mu}$ et $n(n+1)/2$ pour $\boldsymbol{V}$.

Puisque $\boldsymbol{V}$ est définie positive on a $\det \boldsymbol{V} > 0$ et $\boldsymbol{V}^{-1}$ existe, l'expression (6.57) a donc un sens. La matrice $\boldsymbol{V}^{-1}$ est également définie positive et l'expression $Q(\boldsymbol{x}) = (\boldsymbol{x}-\boldsymbol{\mu})^t\boldsymbol{V}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$ représente alors une forme quadratique définie positive homogène en $\boldsymbol{x}-\boldsymbol{\mu}$. L'équation $Q(\boldsymbol{x}) = k^2$ est celle d'un ellipsoïde, l'ellipsoïde $Q(\boldsymbol{x}) = 1$ est dit : ellipsoïde de corrélation.

On note $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{V})$ un vecteur aléatoire qui suit une loi normale nD de paramètres $\boldsymbol{\mu}$ et $\boldsymbol{V}$. On démontrera au paragraphe 6.3.5 que $\boldsymbol{\mu}$ et $\boldsymbol{V}$ sont respectivement la moyenne et la matrice des variances-covariances du vecteur $\boldsymbol{X}$.

6.3.1 Fonction caractéristique nD.

La fonction caractéristique $Z(\boldsymbol{\omega})$ est l'espérance des variables aléatoires complexes $e^{i\boldsymbol{\omega}^t\boldsymbol{X}}$. Il vient :

$$Z(\boldsymbol{\omega}) = \exp\{-\tfrac{1}{2}\boldsymbol{\omega}^t\boldsymbol{V}\boldsymbol{\omega}\}\exp\{i\boldsymbol{\omega}^t\boldsymbol{\mu}\}\,. \tag{6.58}$$

6.3.2 Changement de variable linéaire.

Nous allons montrer qu'une combinaison linéaire non-singulière des variables aléatoires normales $\boldsymbol{X}$ reste normale. Plus précisément, nous avons le théorème suivant :

Théorème 6.1. *Soit $\boldsymbol{X}$ une variable aléatoire normale à n dimensions de moyenne $\boldsymbol{\mu}$ et de matrice des variances-covariances $\boldsymbol{V}$. Soit un changement de variable linéaire : $\boldsymbol{Y} = \boldsymbol{B}\boldsymbol{X}$, où $\boldsymbol{B}$ est une matrice carrée régulière.*

Les nouvelles variables aléatoires sont alors normales à n dimensions, de moyenne $\boldsymbol{B}\boldsymbol{\mu}$ et de matrice des variances-covariances $\boldsymbol{B}\boldsymbol{V}\boldsymbol{B}^t$. Soit :

$$[\,\boldsymbol{X} = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{V}),\ \det\boldsymbol{B} \neq 0\,] \Longrightarrow [\,\boldsymbol{B}\boldsymbol{X} = \mathcal{N}(\boldsymbol{B}\boldsymbol{\mu}, \boldsymbol{B}\boldsymbol{V}\boldsymbol{B}^t)\,]\,. \tag{6.59}$$

Démonstration. Nous savons, d'après les résultats établis en 5.6.5 dans le cadre des changements de variables linéaires, que $\boldsymbol{Y}$ possède une moyenne $\boldsymbol{B}\boldsymbol{X}$ et une matrice des variances-covariances $\boldsymbol{B}\boldsymbol{V}\boldsymbol{B}^t$. Le seul élément nouveau est que la variable $\boldsymbol{Y}$ reste normale.

La matrice $\boldsymbol{B}$ étant régulière, $\boldsymbol{B}^{-1}$ existe et le changement de variables est bijectif. On obtient alors la densité $f_{\boldsymbol{Y}}$ de $\boldsymbol{Y}$ à partir de celle de $\boldsymbol{X}$ à l'aide de la formule : $f_{\boldsymbol{Y}}(\boldsymbol{y}) = f_{\boldsymbol{X}}(\boldsymbol{x})|J|^{-1}$, où J est le jacobien du changement de variables (voir équation (4.20), page 52). Ce changement de variables étant linéaire, on a $J = \det\boldsymbol{B}$. La forme quadratique $(\boldsymbol{x}-\boldsymbol{\mu})^t\boldsymbol{V}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$ devient $(\boldsymbol{B}^{-1}\boldsymbol{y}-\boldsymbol{\mu})^t\boldsymbol{V}^{-1}(\boldsymbol{B}^{-1}\boldsymbol{y}-\boldsymbol{\mu})$ et en posant $\boldsymbol{\mu}' = \boldsymbol{B}\boldsymbol{\mu}$ il vient $(\boldsymbol{y}-\boldsymbol{\mu}')^t(\boldsymbol{B}^{-1})^t\boldsymbol{V}^{-1}\boldsymbol{B}^{-1}(\boldsymbol{y}-\boldsymbol{\mu}') = (\boldsymbol{y}-\boldsymbol{\mu}')^t(\boldsymbol{B}\boldsymbol{V}\boldsymbol{B}^t)^{-1}(\boldsymbol{y}-\boldsymbol{\mu}')$. Ce calcul montre que la forme quadratique en $\boldsymbol{x}$ reste une forme quadratique en $\boldsymbol{y}$ et qu'ainsi la loi suivie par $\boldsymbol{Y}$ est normale. □

On vérifie que la constante devant la densité de la loi normale suivie par $\boldsymbol{Y}$ est divisée par $|\det\boldsymbol{B}|$ de sorte que sa densité s'écrit bien :

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = \frac{1}{(2\pi)^{\frac{n}{2}}(\det\boldsymbol{V}')^{\frac{1}{2}}}\,\exp\{-\tfrac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu}')^t\boldsymbol{V}'^{-1}(\boldsymbol{y}-\boldsymbol{\mu}')\}\,, \tag{6.60}$$

avec $\boldsymbol{\mu}' = \boldsymbol{B}\boldsymbol{\mu}$ et $\boldsymbol{V}' = \boldsymbol{B}\boldsymbol{V}\boldsymbol{B}^t$. La table 6.2 donne les caractéristiques numériques de certains changements de variables linéaires.

▶**Exemple 6.1.** *Somme et différence de deux variables aléatoires normales corrélées.* Soit $\boldsymbol{X}$ une variable aléatoire suivant la loi normale 2D de moyenne $\boldsymbol{\mu}$ et de matrice des variances-covariances $\boldsymbol{V}$ donnée par l'expression (6.17). On cherche la loi suivie par la nouvelle variable aléatoire $\boldsymbol{Y} = \boldsymbol{B}\boldsymbol{X}$, de composantes Y_1, Y_2 telles que :

$$\begin{aligned} Y_1 &= X_1 + X_2\,, \\ Y_2 &= X_1 - X_2\,. \end{aligned} \quad \text{On a alors} \quad \boldsymbol{B} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}\,.$$

Variable	$\boldsymbol{X}$	$\boldsymbol{X}$-$\boldsymbol{\mu}$	$\boldsymbol{BX}$	$\boldsymbol{U}^t(\boldsymbol{X}$-$\boldsymbol{\mu})$	$\boldsymbol{\Lambda}^{-1}\boldsymbol{U}^t(\boldsymbol{X}$-$\boldsymbol{\mu})$
Moyenne	$\boldsymbol{\mu}$	$\mathbf{0}$	$\boldsymbol{B\mu}$	$\mathbf{0}$	$\mathbf{0}$
Variances-Covariances	$\boldsymbol{V}$	$\boldsymbol{V}$	$\boldsymbol{BVB}^t$	$\boldsymbol{\Lambda}^2$	$\boldsymbol{I}$

TAB. 6.2: *Caractéristiques numériques de certains changements de variable où la variable aléatoire $\boldsymbol{X}$ suit une loi normale à n dimensions (nD). La matrice unitaire $\boldsymbol{U}$ est la matrice de changement de base qui diagonalise $\boldsymbol{V}$, elle est telle que $\boldsymbol{\Lambda}^2 = \boldsymbol{U}^t\boldsymbol{V}\boldsymbol{U}$. La matrice $\boldsymbol{I}$ est la matrice identité et $\mathbf{0}$ la matrice nulle. Les nouvelles variables aléatoires définies par ces transformations suivent toutes la loi normale nD.*

D'après les résultats ci-dessus, la nouvelle variable aléatoire $\boldsymbol{Y}$ suit une loi normale (2D) de moyenne $\boldsymbol{\mu'} = \boldsymbol{B\mu}$ et de matrice des variances-covariances $\boldsymbol{V'} = \boldsymbol{BVB}^t$ (voir table 6.2). On a pour la moyenne :

$$\boldsymbol{\mu'} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} = \begin{pmatrix} \mu_1 + \mu_2 \\ \mu_1 - \mu_2 \end{pmatrix} , \tag{6.61}$$

et pour la nouvelle matrice des variances-covariances $\boldsymbol{V'}$:

$$\begin{aligned} \boldsymbol{V'} &= \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} . \\ \boldsymbol{V'} &= \begin{pmatrix} \sigma_1^2 + 2\rho\sigma_1\sigma_2 + \sigma_2^2 & \sigma_1^2 - \sigma_2^2 \\ \sigma_1^2 - \sigma_2^2 & \sigma_1^2 - 2\rho\sigma_1\sigma_2 + \sigma_2^2 \end{pmatrix} . \end{aligned} \tag{6.62}$$

Les lois suivies par $X_1 + X_2$ et $X_1 - X_2$ sont les lois marginales de $\boldsymbol{Y}$, d'après le résultat du paragraphe 6.2.7 ce sont des lois normales. La table 6.3 donne l'expression de la moyenne et de la variance de la somme et de la différence de deux variables aléatoires normales corrélées.

Variable	X_1	X_2	$X_1 + X_2$	$X_1 - X_2$
Moyenne	μ_1	μ_2	$\mu_1 + \mu_2$	$\mu_1 - \mu_2$
Variance	σ_1^2	σ_2^2	$\sigma_1^2 + 2\rho\sigma_1\sigma_2 + \sigma_2^2$	$\sigma_1^2 - 2\rho\sigma_1\sigma_2 + \sigma_2^2$

TAB. 6.3: *Moyenne et variance de la somme et de la différence de deux variables aléatoires normales corrélées. Cette somme et cette différence suivent des lois normales. On retrouve le résultat classique dans le cas $\rho = 0$.*

6.3.3 Loi normale nD réduite.

Un vecteur aléatoire $\boldsymbol{X}$ suit la loi normale nD *réduite* si, dans l'expression (6.57) précédente, $\boldsymbol{\mu}$ est nul et $\boldsymbol{V}$ est la matrice identité. Sa densité de probabilité s'écrit alors :

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{n}{2}}} \exp\{-\tfrac{1}{2}\sum_{i=1}^{n} x_i^2\} = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\{-\tfrac{1}{2}x_i^2\} . \tag{6.63}$$

La dernière égalité montre que les composantes X_i du vecteur $\boldsymbol{X}$ suivent la loi normale réduite et sont mutuellement *indépendantes*.

6.3.4 Réduction des variables normales quelconques.

Il est toujours possible de transformer un vecteur aléatoire normal $\boldsymbol{X}$ quelconque en un vecteur $\boldsymbol{Y}$ normal réduit. Il suffit pour cela de réduire la forme quadratique Q à sa forme diagonale.

Théorème 6.2. *Soit $\boldsymbol{X}$ un vecteur aléatoire normal de paramètres $\boldsymbol{\mu}$ et $\boldsymbol{V}$. Il existe alors une matrice diagonale $\boldsymbol{\Lambda}$ et une matrice unitaire $\boldsymbol{U}$ telles que le vecteur aléatoire $\boldsymbol{Y}$ défini par le changement de variable :*

$$\boldsymbol{X} - \boldsymbol{\mu} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{Y}\,, \tag{6.64}$$

suit la loi normale nD réduite. La matrice $\boldsymbol{\Lambda}^2$ est la matrice des valeurs propres de $\boldsymbol{V}$ et la matrice $\boldsymbol{U}$ est la matrice unitaire de ses vecteurs propres, $\boldsymbol{\Lambda}$ et $\boldsymbol{U}$ satisfont donc les relations :

$$\boldsymbol{U}^{-1} = \boldsymbol{U}^t, \quad \boldsymbol{\Lambda}^2 = \boldsymbol{U}^t\boldsymbol{V}\boldsymbol{U}\,. \tag{6.65}$$

Démonstration. La densité de probabilité du vecteur $\boldsymbol{X} - \boldsymbol{\mu}$ est symétrique en $\boldsymbol{x} - \boldsymbol{\mu}$, elle est donc de moyenne nulle, soit : $\mathrm{E}\{\boldsymbol{X} - \boldsymbol{\mu}\} = \boldsymbol{0}$. On passe ensuite des variables $\boldsymbol{X} - \boldsymbol{\mu}$ aux variables $\boldsymbol{Y}$ par le changement de variable linéaire de matrice : $\boldsymbol{B} = \boldsymbol{\Lambda}^{-1}\boldsymbol{U}^t$, la matrice $\boldsymbol{V}$ étant définie positive ses valeurs propres sont strictement positives et par conséquent $\boldsymbol{\Lambda}^{-1}$ existe.

D'après le théorème 6.1, le vecteur $\boldsymbol{Y}$ suit une loi normale de moyenne : $\mathrm{E}\{\boldsymbol{Y}\} = \boldsymbol{\Lambda}^{-1}\boldsymbol{U}^t\,\mathrm{E}\{\boldsymbol{X} - \boldsymbol{\mu}\} = \boldsymbol{0}$ et de matrice des variances-covariances : $\boldsymbol{\Lambda}^{-1}\boldsymbol{U}^t\boldsymbol{V}(\boldsymbol{\Lambda}^{-1}\boldsymbol{U}^t)^t = \boldsymbol{\Lambda}^{-1}\boldsymbol{U}^t\boldsymbol{V}\boldsymbol{U}\boldsymbol{\Lambda}^{-1}$. Par hypothèse on a les expressions (6.65) et la matrice des variances--covariances de $\boldsymbol{Y}$ s'écrit $\boldsymbol{\Lambda}^{-1}\boldsymbol{\Lambda}^2\boldsymbol{\Lambda}^{-1} = \boldsymbol{I}$. Ce dernier point démontre que la variable aléatoire $\boldsymbol{Y}$ suit effectivement une loi normale réduite. □

Le changement de variable s'écrit explicitement :

$$\boldsymbol{y} = \boldsymbol{\Lambda}^{-1}\boldsymbol{U}^t(\boldsymbol{x} - \boldsymbol{\mu})\,, \tag{6.66}$$

il exprime que pour obtenir les variables $\boldsymbol{y}$ à partir des variables $\boldsymbol{x}$, il faut d'abord procéder à une translation puis à une rotation et enfin à un changement d'échelle. Au cours de ces transformations l'ellipse de corrélation devient un cercle de rayon unité centré sur l'origine des axes.

6.3.5 Caractéristiques numériques de la loi normale à plusieurs variables.

Moyenne et matrice des variances-covariances. Le théorème précédent permet de trouver la signification des paramètres $\boldsymbol{\mu}$ et $\boldsymbol{V}$ entrant dans l'expression de la densité de probabilité du vecteur aléatoire $\boldsymbol{X}$. Ce sont respectivement la moyenne et la matrice des variances-covariances de $\boldsymbol{X}$, on a :

$$\mathrm{E}\{\boldsymbol{X}\} = \boldsymbol{\mu}\,, \tag{6.67}$$

$$\mathrm{E}\{(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^t\} = \boldsymbol{V}\,. \tag{6.68}$$

Démonstration. D'après le théorème 6.2 on a $\mathrm{E}\{\boldsymbol{Y}\} = \boldsymbol{0}$ et $\mathrm{E}\{\boldsymbol{Y}\boldsymbol{Y}^t\} = \boldsymbol{I}$, il vient pour la moyenne : $\mathrm{E}\{\boldsymbol{X} - \boldsymbol{\mu}\} = \mathrm{E}\{\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{Y}\} = \boldsymbol{U}\boldsymbol{\Lambda}\,\mathrm{E}\{\boldsymbol{Y}\} = \boldsymbol{0}$ d'où le premier résultat. Pour la matrice des variances-covariances, par définition elle est égale à : $\mathrm{E}\{(\boldsymbol{X} -$

$\mathrm{E}\{\boldsymbol{X}\})(\boldsymbol{X}-\mathrm{E}\{\boldsymbol{X}\})^t\} = \mathrm{E}\{(\boldsymbol{X}-\boldsymbol{\mu})(\boldsymbol{X}-\boldsymbol{\mu})^t\} = \mathrm{E}\{\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{Y}\boldsymbol{Y}^t\boldsymbol{\Lambda}\boldsymbol{U}^t\} = \boldsymbol{U}\boldsymbol{\Lambda}\mathrm{E}\{\boldsymbol{Y}\boldsymbol{Y}^t\}\boldsymbol{\Lambda}\boldsymbol{U}^t$ d'où $\mathrm{E}\{(\boldsymbol{X}-\mathrm{E}\{\boldsymbol{X}\})(\boldsymbol{X}-\mathrm{E}\{\boldsymbol{X}\})^t\} = \boldsymbol{U}\boldsymbol{\Lambda}^2\boldsymbol{U}^t = \boldsymbol{V}$. La matrice $\boldsymbol{V}$ est bien, comme nous l'avions annoncé, la matrice des variances-covariances du vecteur aléatoire $\boldsymbol{X}$. □

Ainsi la loi normale à n dimensions est entièrement déterminée par la donnée de ses moments jusqu'à l'ordre deux. Les éléments μ_i du vecteur $\boldsymbol{\mu}$ sont les moyennes des lois marginales suivie par les variables X_i. Les éléments $\rho_{ij}\sigma_i\sigma_j$ de la matrice $\boldsymbol{V}$ sont les covariances des couples (X_i, X_j), les ρ_{ij} en sont les coefficients de corrélation et les σ_i les écarts types. On a :

$$\boldsymbol{V} = \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \cdots & \rho_{1n}\sigma_1\sigma_n \\ \rho_{12}\sigma_2\sigma_1 & \sigma_2^2 & \cdots & \rho_{2n}\sigma_2\sigma_n \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1n}\sigma_n\sigma_1^2 & \rho_{2n}\sigma_2\sigma_n & \cdots & \sigma_n^2 \end{pmatrix} . \tag{6.69}$$

6.3.6 Lois marginales et conditionnelles.

Les lois marginales et conditionnelles distinguent certaines composantes du vecteur aléatoire $\boldsymbol{X}$, en nombre r, dites variables « actives » par rapport aux $n-r$ restantes dites « inactives ». Rappelons que pour obtenir les lois marginales on intègre sur les variables inactives alors qu'on les considère constantes pour obtenir les lois conditionnelles.

Afin de simplifier l'exposé, nous supposons que les variables actives sont les r premières composantes du vecteur $\boldsymbol{X}$. Ce vecteur peut alors être partitionné suivant le schéma :

$$(\underbrace{X_1, \ldots, X_r}_{\boldsymbol{X}_0}, \underbrace{X_{r+1}, \ldots, X_n}_{\boldsymbol{X}_1}) . \tag{6.70}$$

A cette partition correspond une partition équivalente des valeurs $\boldsymbol{x}$ prises par $\boldsymbol{X}$ et de la moyenne $\boldsymbol{\mu}$ respectivement en $(\boldsymbol{x}_0, \boldsymbol{x}_1)$ et $(\boldsymbol{\mu}_0, \boldsymbol{\mu}_1)$. La matrice des variances-covariances et son inverse se partionnent en 4 matrices blocs :

$$\boldsymbol{V} = \begin{pmatrix} \boldsymbol{V}_{00} & \boldsymbol{V}_{01} \\ \boldsymbol{V}_{10} & \boldsymbol{V}_{11} \end{pmatrix} , \quad \boldsymbol{V}^{-1} = \begin{pmatrix} \boldsymbol{A}_{00} & \boldsymbol{A}_{01} \\ \boldsymbol{A}_{10} & \boldsymbol{A}_{11} \end{pmatrix} .$$

Les matrices $\boldsymbol{V}_{00}$ et $\boldsymbol{V}_{11}$ sont carrées et symétriques respectivement de format (r, r) et $(n-r, n-r)$. Les matrices $\boldsymbol{V}_{01}$ et $\boldsymbol{V}_{10}$ sont rectangulaires, de format $(r, n-r)$ et $(n-r, r)$, on a $\boldsymbol{V}_{10} = \boldsymbol{V}_{01}^t$. Il existe des propriétés analogues pour les matrices blocs composant $\boldsymbol{V}^{-1}$.

Lois marginales.

Moments. Les moments des variables aléatoires $\boldsymbol{X}_0$ sont les mêmes, qu'ils soient calculés suivant la loi nD ou suivant la loi marginale correspondante, ceci signifie que la moyenne de $\boldsymbol{X}_0$ est égale à $\boldsymbol{\mu}_0$ et que sa matrice des variances-covariances est égale à $\boldsymbol{V}_{00}$. En particulier on a :

$$\mathrm{E}\{X_i\} = \mu_i, \quad \mathrm{Var}(X_i) = \sigma_i^2, \quad \mathrm{Cov}(X_i, X_j) = \rho_{ij}\sigma_i\sigma_j \tag{6.71}$$

Densité de probabilité. Nous allons maintenant montrer que les lois marginales suivent aussi une loi normale. Dans ce but considérons la forme quadratique $Q = (\boldsymbol{x} - \boldsymbol{\mu})^t \boldsymbol{V}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})$ entrant dans l'expression de la loi normale. Eliminons les moyennes à l'aide du changement de variable bijectif $\boldsymbol{y} = \boldsymbol{x} - \boldsymbol{\mu}$ et développons Q suivant la partition. Il vient :

$$Q = \boldsymbol{y}_0^t \boldsymbol{A}_{00} \boldsymbol{y}_0 + 2\boldsymbol{y}_0^t \boldsymbol{A}_{01} \boldsymbol{y}_1 + \boldsymbol{y}_1^t \boldsymbol{A}_{11} \boldsymbol{y}_1 \,. \tag{6.72}$$

Effectuons, afin d'éliminer le terme croisé $\boldsymbol{y}_0^t \boldsymbol{A}_{01} \boldsymbol{y}_1$, un deuxième changement de variable, lui aussi bijectif : $\boldsymbol{z}_1 = \boldsymbol{y}_1 + \boldsymbol{a}$. Il est immédiat de déterminer que la constante $\boldsymbol{a}$ qui réalise cette élimination est $\boldsymbol{a} = \boldsymbol{A}_{11}^{-1} \boldsymbol{A}_{10} \boldsymbol{y}_0$. On trouve alors :

$$Q = \boldsymbol{y}_0^t (\boldsymbol{A}_{00} - \boldsymbol{A}_{01} \boldsymbol{A}_{11}^{-1} \boldsymbol{A}_{10}) \boldsymbol{y}_0 + \boldsymbol{z}_1^t \boldsymbol{A}_{11} \boldsymbol{z}_1 \equiv Q_0 + Q_1 \,. \tag{6.73}$$

Nous venons donc de montrer que la loi nD pouvait se mettre sous la forme du produit de deux lois normales, une rD par une $(n-r)$D. On obtient la loi marginale en intégrant sur les variables $\boldsymbol{y}_1$ ou, ce qui revient au même, sur les variables $\boldsymbol{z}_1$. La deuxième loi normale donne 1 par intégration, de sorte qu'il ne reste plus dans l'expression de la loi marginale que le premier terme Q_0, ce qui montre que la loi marginale est normale. Par ailleurs nous savons que sa matrice des variances-covariances est égale à $\boldsymbol{V}_{00}$, ce qui implique nécessairement que :

$$\boldsymbol{V}_{00}^{-1} = \boldsymbol{A}_{00} - \boldsymbol{A}_{01} \boldsymbol{A}_{11}^{-1} \boldsymbol{A}_{10}. \tag{6.74}$$

En revenant aux variables initiales, on trouve la densité de probabilité de la loi marginale suivie par $\boldsymbol{X}_0$:

$$f_{\boldsymbol{X}_0}(\boldsymbol{x}_0) = \frac{1}{(2\pi)^{\frac{1}{2}\dim \boldsymbol{X}_0} (\det \boldsymbol{V}_{00})^{\frac{1}{2}}} \exp\{-\tfrac{1}{2}(\boldsymbol{x}_0 - \boldsymbol{\mu}_0)^t \boldsymbol{V}_{00}^{-1} (\boldsymbol{x}_0 - \boldsymbol{\mu}_0)\} \,, \tag{6.75}$$

où $\dim \boldsymbol{X}_0$ est ègal à la dimension du vecteur $\boldsymbol{X}_0$ (ici $\dim \boldsymbol{X}_0 = r$).

Lois conditionnelles.

Afin de déterminer, par exemple, la densité conditionnelle de $\boldsymbol{X}_1$ connaissant $\boldsymbol{X}_0$: $f_{\boldsymbol{X}_1|\boldsymbol{X}_0}$, nous allons de nouveau utiliser la relation qui lie la densité de $\boldsymbol{X}$ avec les densités marginales et conditionnelles. Cette relation s'écrit :

$$f_{\boldsymbol{X}_0 \boldsymbol{X}_1}(\boldsymbol{x}_0, \boldsymbol{x}_1) = f_{\boldsymbol{X}_0}(\boldsymbol{x}_0) f_{\boldsymbol{X}_1|\boldsymbol{X}_0}(\boldsymbol{x}_1 | \boldsymbol{X}_0 = \boldsymbol{x}_0) \,. \tag{6.76}$$

Dans cette formule, nous avons noté $f_{\boldsymbol{X}_0 \boldsymbol{X}_1}$ la densité de probabilité du vecteur $\boldsymbol{X}$ qui, d'après (6.73) et (6.74), peut s'écrire :

$$f_{\boldsymbol{X}_0 \boldsymbol{X}_1}(\boldsymbol{x}_0, \boldsymbol{x}_1) = \frac{1}{(2\pi)^{\frac{n}{2}} (\det \boldsymbol{V})^{\frac{1}{2}}} \exp\{-\tfrac{1}{2} \boldsymbol{y}_0{}^t \boldsymbol{V}_{00}^{-1} \boldsymbol{y}_0 + \boldsymbol{z}_1^t \boldsymbol{A}_{11} \boldsymbol{z}_1\}, \tag{6.77}$$

avec $\boldsymbol{y}_0 = \boldsymbol{x}_0 - \boldsymbol{\mu}_0$, $\boldsymbol{y}_1 = \boldsymbol{x}_1 - \boldsymbol{\mu}_1$ et $\boldsymbol{z}_1 = \boldsymbol{y}_1 + \boldsymbol{A}_{11}^{-1} \boldsymbol{A}_{10} \boldsymbol{y}_0$. En divisant la densité nD (6.77) par la densité marginale (6.75) on obtient la densité conditionnelle cherchée. Après simplification des termes de l'exponentielle, il ne reste que la forme quadratique $Q_1 = \boldsymbol{z}_1^t \boldsymbol{A}_{11} \boldsymbol{z}_1$, on en déduit l'expression de la densité conditionnelle :

$$f_{\boldsymbol{X}_1|\boldsymbol{X}_0}(\boldsymbol{x}_1 | \boldsymbol{X}_0 = \boldsymbol{x}_0) = \frac{(\det \boldsymbol{A}_{11})^{\frac{1}{2}}}{(2\pi)^{\frac{n-r}{2}}} \exp\{-\tfrac{1}{2} \boldsymbol{z}_1^t \boldsymbol{A}_{11} \boldsymbol{z}_1\} \,. \tag{6.78}$$

En revenant aux variables initiales la forme quadratique Q_1 s'écrit :

$$Q_1(\boldsymbol{x}_1) = (\boldsymbol{x}_1 - \boldsymbol{\mu}_1 + \boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{10}\boldsymbol{y}_0)^t \boldsymbol{A}_{11}(\boldsymbol{x}_1 - \boldsymbol{\mu}_1 + \boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{10}\boldsymbol{y}_0)\,, \tag{6.79}$$

ce qui montre que la densité conditionnelle $f_{\boldsymbol{X}_1|\boldsymbol{X}_0}$ est celle d'une variable aléatoire normale de moyenne $\boldsymbol{\mu}_1 - \boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{10}\boldsymbol{y}_0$ et de matrice des variances-covariances $\boldsymbol{A}_{11}^{-1}$.

En interchangeant les indices 0 et 1 et les symboles $\boldsymbol{A}$ et $\boldsymbol{V}$ dans (6.74) on trouve que : $\boldsymbol{A}_{11}^{-1} = \boldsymbol{V}_{11} - \boldsymbol{V}_{10}\boldsymbol{V}_{00}^{-1}\boldsymbol{V}_{01}$. En exprimant que $\boldsymbol{V}\boldsymbol{V}^{-1} = \boldsymbol{V}^{-1}\boldsymbol{V} = \boldsymbol{I}$ avec les blocs, on trouve que $\boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{10} = -\boldsymbol{V}_{10}\boldsymbol{V}_{00}^{-1}$. Ce qui permet d'exprimer la moyenne et la matrice des variances-covariances de la loi conditionnelle en fonction de la moyenne et la matrice des variances-covariances de la loi à n dimensions. Ces expressions sont données ci-dessous.

Moyenne et variances conditionnelles. Dans le paragraphe précédent, nous avons obtenu la moyenne et la matrice des variances-covariances de la loi conditionnelle de $\boldsymbol{X}_1$ sachant que $\boldsymbol{X}_0 = \boldsymbol{x}_0$. On a pour la moyenne conditionnelle :

$$\mathrm{E}\{\boldsymbol{X}_1|\boldsymbol{X}_0 = \boldsymbol{x}_0\} = \boldsymbol{\mu}_1 + \boldsymbol{V}_{10}\boldsymbol{V}_{00}^{-1}(\boldsymbol{x}_0 - \boldsymbol{\mu}_0)\,. \tag{6.80}$$

Soit $\boldsymbol{\mu}_{1|0}$ cette valeur, il vient pour la matrice des variances-covariances conditionnelle :

$$\mathrm{E}\{(\boldsymbol{X}_1 - \boldsymbol{\mu}_{1|0})(\boldsymbol{X}_1 - \boldsymbol{\mu}_{1|0})^t|\boldsymbol{X}_0 = \boldsymbol{x}_0\} = \boldsymbol{V}_{11} - \boldsymbol{V}_{10}\boldsymbol{V}_{00}^{-1}\boldsymbol{V}_{01}\,. \tag{6.81}$$

Par permutation des indices 0 et 1, on obtient pour la loi conditionnelle de $\boldsymbol{X}_0$ sachant que $\boldsymbol{X}_1 = \boldsymbol{x}_1$:

$$\mathrm{E}\{\boldsymbol{X}_0|\boldsymbol{X}_1 = \boldsymbol{x}_1\} = \boldsymbol{\mu}_0 + \boldsymbol{V}_{01}\boldsymbol{V}_{11}^{-1}(\boldsymbol{x}_1 - \boldsymbol{\mu}_1)\,, \tag{6.82}$$

$$\mathrm{E}\{(\boldsymbol{X}_0 - \boldsymbol{\mu}_{0|1})(\boldsymbol{X}_0 - \boldsymbol{\mu}_{0|1})^t|\boldsymbol{X}_1 = \boldsymbol{x}_1\} = \boldsymbol{V}_{00} - \boldsymbol{V}_{01}\boldsymbol{V}_{11}^{-1}\boldsymbol{V}_{10}\,. \tag{6.83}$$

Densité de probabilité conditionelles. Les densités conditionelles étant normales, elles sont entièrement déterminées par la donnée de leurs moyennes et de leurs matrices des variances-covariances. Ces quantités peuvent être trouvées ci-dessus, il suffit alors de préciser la constante de normalisation que nous donnons en fonction de la matrice $\boldsymbol{V}$ et de ces blocs associés. Il vient pour la densité de $\boldsymbol{X}_1$ connaissant $\boldsymbol{X}_0$:

$$[\,f_{\boldsymbol{X}_1|\boldsymbol{X}_0}(\boldsymbol{x}_1|\boldsymbol{x}_0)\,] : \frac{1}{(2\pi)^{\frac{1}{2}\dim \boldsymbol{V}_{11}}}\left[\frac{\det \boldsymbol{V}_{00}}{\det \boldsymbol{V}}\right]^{\frac{1}{2}}\,, \tag{6.84}$$

et pour celle de $\boldsymbol{X}_0$ connaissant $\boldsymbol{X}_1$:

$$[\,f_{\boldsymbol{X}_0|\boldsymbol{X}_1}(\boldsymbol{x}_0|\boldsymbol{x}_1)\,] : \frac{1}{(2\pi)^{\frac{1}{2}\dim \boldsymbol{V}_{00}}}\left[\frac{\det \boldsymbol{V}_{11}}{\det \boldsymbol{V}}\right]^{\frac{1}{2}}\,. \tag{6.85}$$

►**Exemple 6.2.** *Une seule variable fixée.* Soit $\boldsymbol{X}$ un vecteur aléatoire normal de moyenne $\boldsymbol{\mu} = 0$ et de matrice des variances-covariances $\boldsymbol{V}$. On demande la moyenne et la matrice des variances-covariances de la loi conditionnelle de $\boldsymbol{X}$ lorsqu'une variable est connue.

Supposons que la variable connue soit la dernière composante de $\boldsymbol{X}$. Réalisons une partition de $\boldsymbol{X}$ suivant le schéma (6.70), on a $\boldsymbol{X}_1 = X_n$ et il vient :

$$\boldsymbol{\mu}_0 = \mathbf{0}\,, \quad \boldsymbol{\mu}_1 = 0\,,$$

$$\begin{pmatrix} \boldsymbol{V}_{00} & \boldsymbol{V}_{01} \\ \boldsymbol{V}_{10} & \boldsymbol{V}_{11} \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \cdots & \rho_{1,n-1}\sigma_1\sigma_{n-1} & \rho_{1n}\sigma_1\sigma_n \\ \vdots & \ddots & \vdots & \vdots \\ \rho_{n-1,1}\sigma_{n-1}\sigma_1 & \cdots & \sigma_n^2 & \rho_{n-1,n}\sigma_{n-1}\sigma_n \\ \rho_{n1}\sigma_n\sigma_1 & \cdots & \rho_{n,n-1}\sigma_n\sigma_{n-1} & \sigma_n^2 \end{pmatrix}.$$

En appliquant les résultats (6.82) et (6.83), on trouve la moyenne conditionelle : $\boldsymbol{\mu}'$ et la matrice des variances-covariances conditionelle : $\boldsymbol{V}'$ du vecteur $\boldsymbol{X}_0$ sachant que $\boldsymbol{X}_1 = x_n$:

$$\boldsymbol{\mu}' = \boldsymbol{V}_{01}\boldsymbol{V}_{11}^{-1}x_n = \begin{pmatrix} \rho_{1n}\dfrac{\sigma_1}{\sigma_n} \\ \rho_{2n}\dfrac{\sigma_2}{\sigma_n} \\ \vdots \\ \rho_{n-1,n}\dfrac{\sigma_{n-1}}{\sigma_n} \end{pmatrix} x_n\,, \tag{6.86}$$

$$\begin{aligned} \boldsymbol{V}' &= \boldsymbol{V}_{00} - \boldsymbol{V}_{01}\boldsymbol{V}_{11}^{-1}\boldsymbol{V}_{10} \\ &= \begin{pmatrix} \sigma_1^2(1-\rho_{1n}^2) & \cdots & \sigma_1\sigma_{n-1}(\rho_{1,n-1} - \rho_{1n}\rho_{n,n-1}) \\ \vdots & \ddots & \vdots \\ \sigma_{n-1}\sigma_1(\rho_{n-1,1} - \rho_{n-1,n}\rho_{n1}) & \cdots & \sigma_{n-1}^2(1-\rho_{n-1,n}^2) \end{pmatrix}. \end{aligned} \tag{6.87}$$

On constate alors que l'effet du conditionnement est de modifier les covariances et les coefficients de corrélations entre les variables restées actives. En particulier l'écart type des variables actives est réduit du fait de leur corrélation éventuelle avec la variable connue (inactive). Si $\rho_{ij}^{(n)}$ désigne le coefficient de corrélation des variables X_i et X_j connaissant X_n, il vient :

$$\rho_{ij}^{(n)} = \frac{\rho_{ij} - \rho_{in}\rho_{jn}}{[1-\rho_{in}^2]^{\frac{1}{2}}[1-\rho_{jn}^2]^{\frac{1}{2}}}\,. \tag{6.88}$$

Si l'on pose $x_n = \sigma_n^2$ dans l'équation (6.86) on obtient $\boldsymbol{\mu}' = \boldsymbol{V}_{01}$ ce qui veut dire qu'en ce point les coordonnées de la moyenne conditionelle $\boldsymbol{\mu}'$ dans l'espace $\mathbb{R}^n$ où se répartit $\boldsymbol{X}$ est identique à la n^{e} colonne de la matrice des variances-covariances $\boldsymbol{V}$ de $\boldsymbol{X}$.

6.3.7 Ellipsoïde d'égale densité.

Rappelons que le lieu des points où $Q(\boldsymbol{x}) = k^2$ est un ellipsoïde. Sur cet ellipsoïde la densité de probabilité $f(\boldsymbol{x})$ de la loi normale nD est constante. Cet ellipsoïde est appelé l'ellipsoïde d'égale densité de probabilité ou plus simplement l'*ellipsoïde d'égale densité*. Nous avons déjà mentionné que si l'on pose $k^2 = 1$ cet ellipsoïde prend alors le nom d'*ellipsoïde de corrélation*.

On peut également montrer que l'hyper-rectangle (de dispersion) parallèle aux axes x_i et circonscrit à l'ellipsoïde de corrélation, possède des arêtes de longueurs $2\sigma_i$. Nous énonçons cette propriété sous une forme équivalente dans le théorème qui suit.

Théorème 6.3. *Soit $\boldsymbol{X}$ un vecteur aléatoire normal à n dimensions de moyenne $\boldsymbol{\mu}$ et de matrice des variances-covariances $\boldsymbol{V}$. Les valeurs extrêmes atteintes par*

les coordonnées $\boldsymbol{x}$ de l'ellipsoïde d'égale probabilité d'équation : $Q(\boldsymbol{x}) = k^2$ sont égales à $\mu_i \pm k\sigma_i$. Plus précisément on a :

$$\min_{x_i}\{x_i | Q(\boldsymbol{x}) = k^2\} = \mu_i - k\sigma_i\,, \quad \max_{x_i}\{x_i | Q(\boldsymbol{x}) = k^2\} = \mu_i + k\sigma_i\,. \tag{6.89}$$

On notera que ces valeurs extrêmes sont indépendantes des coefficients de corrélations.

Démonstration. Sans nuire à la généralité de la démonstration nous allons supposer que $\boldsymbol{\mu} = \mathbf{0}$ et chercher les extrema de $Q(\boldsymbol{x})$ sur l'axe x_n. Considérons la densité conditionnelle de $\boldsymbol{X}$ sachant que $X_n = k\sigma_n$. Cette densité est celle d'une loi normale de moyenne, dans l'espace $\mathbb{R}^n$, notée $\boldsymbol{\mu}_n$. D'après le résultat de l'exemple 6.2, $\boldsymbol{\mu}_n$ a pour coordonnées :

$$\boldsymbol{\mu}_n = \frac{k}{\sigma_n}\boldsymbol{V}_n\,,$$

où $\boldsymbol{V}_n$ désigne le n^{e} colonne de la matrice $\boldsymbol{V}$. Evaluons la forme quadratique Q au point $\boldsymbol{\mu}_n$, il vient :

$$Q(\boldsymbol{\mu}_n) = \boldsymbol{\mu}_n^t \boldsymbol{V}^{-1}\boldsymbol{\mu}_n = \frac{k^2}{\sigma_n^2}\boldsymbol{V}_n^t\boldsymbol{V}^{-1}\boldsymbol{V}_n = \frac{k^2}{\sigma_n^2}\boldsymbol{V}_n^t \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix} = k^2\,.$$

La densité conditionnelle en tant que densité normale est unimodale et son mode est égal à sa moyenne. Le point $\boldsymbol{\mu}_n$ est alors l'unique point où $Q(\boldsymbol{x}|x_n = k\sigma_n)$ est égal à k^2, il s'agit du point de tangence du plan $x_n = k\sigma_n$ avec l'ellipsoïde $Q(\boldsymbol{x}) = k^2$ et donc du point extrême de cet ellispoïde. En $x_n = k\sigma_n$ on obtient le maximum et en $x_n = -k\sigma_n$ le minimum. □

6.3.8 Composantes principales.

Considérons la variable aléatoire $\boldsymbol{Y}$ définie par le changement de variables aléatoires $(\boldsymbol{X} - \boldsymbol{\mu}) = \boldsymbol{U}\boldsymbol{Y}$, notons $\boldsymbol{x}$ et $\boldsymbol{y}$ les valeurs prises par les anciennes $(\boldsymbol{X})$ et les nouvelles $(\boldsymbol{Y})$ variables aléatoires. On a $(\boldsymbol{x} - \boldsymbol{\mu}) = \boldsymbol{U}\boldsymbol{y}$. De façon triviale, la translation de vecteur $\boldsymbol{\mu}$ ne change que la moyenne de la densité qui devient nulle. Le changement de base de matrice $\boldsymbol{U}$ suivant la discussion de la section 6.3.9 est tel que $\boldsymbol{B} = \boldsymbol{U}^{-1} = \boldsymbol{U}^t$ avec $\det \boldsymbol{U} = 1$. En appliquant les résultats résumés dans la table 6.2 on trouve que la matrice des variances-covariances de $\boldsymbol{Y}$ est égale à $\boldsymbol{U}^t\boldsymbol{V}\boldsymbol{U}$. Cette expression est, d'après (6.65), identique à $\boldsymbol{\Lambda}^2$. La densité de $\boldsymbol{Y}$ est donc égale à :

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = \frac{1}{(2\pi)^{\frac{n}{2}}\det\boldsymbol{\Lambda}}\exp\{-\tfrac{1}{2}\boldsymbol{y}^t\boldsymbol{\Lambda}^{-2}\boldsymbol{y}\}, \quad \boldsymbol{y}^t\boldsymbol{\Lambda}^{-2}\boldsymbol{y} = \sum_{i=1}^{n}\frac{y_i^2}{\lambda_i^2}\,. \tag{6.90}$$

Les composantes Y_i de $\boldsymbol{Y}$ suivent bien une loi normale de moyenne nulle et de variance λ_i^2. Elles sont également non-corrélées et donc indépendantes dans le cas particulier de la loi normale (voir table 6.2). Les vecteurs propres orthonormés rangés en colonne dans $\boldsymbol{U}$ sont appelés les « *composantes principales* » de la loi, ils correspondent aux axes principaux des ellipsoïdes d'égale probabilité. On note $\boldsymbol{u}_i$ ces composantes principales et on les considère en général comme étant

rangées dans l'ordre croissant des valeurs propres λ_i. En écrivant le changement de variable à l'aide des $\boldsymbol{u}_i$ on obtient :

$$\boldsymbol{X} - \boldsymbol{\mu} = \sum_{i=1}^{n} Y_i \boldsymbol{u}_i \,. \tag{6.91}$$

L'expression précédente montre que les composantes principales permettent de décomposer le vecteur aléatoire $\boldsymbol{X}$ sur une base orthonormée non aléatoire formée des vecteurs propres de sa matrice des variances-covariances. Les coefficients Y_i de la décomposition sont des variables aléatoires non-corrélées (et dans le cas normal ils suivent une loi normale et sont donc indépendants), leur moyenne est nulle et leur variance est égale à la valeur propre λ_i^2 qui correspond à la composante principale de même indice.

Cette décomposition en composantes principales ou encore « *canonique* » jouit de nombreuses propriétés remarquables. Par exemple, l'hyper-rectangle de dispersion associé à la base des composantes principales est de volume minimum, voir figure 6.5. Cette démonstration simple repose sur le fait qu'une matrice définie positive (comme $\boldsymbol{V}$) a un déterminant toujours inférieur ou égal au produit de ses éléments diagonaux (voir Bellman 1970, [6]). Cette décomposition possède également plusieurs propriétés optimales vis-à-vis de la troncature de la somme (6.91). Cette décomposition et ses « bonnes » propriétés se généralisent au cas où la loi suivie par $\boldsymbol{X}$ n'est pas normale.

6.3.9 Loi du χ^2.

Un ellipsoïde d'égale densité $Q(\boldsymbol{x}) = k^2$ étant donné, on souhaite calculer la probabilité P_{k^2} pour qu'un point $\boldsymbol{X}$, suivant la loi normale nD, se trouve à l'intérieur de cet ellipsoïde . Le point $\boldsymbol{X}$ sera à l'intérieur de l'ellipsoïde si :

$$\chi^2 = (\boldsymbol{X} - \boldsymbol{\mu})^t \boldsymbol{V}^{-1} (\boldsymbol{X} - \boldsymbol{\mu}) \leq k^2 \,. \tag{6.92}$$

La quantité χ^2 que nous venons d'introduire est une variable aléatoire positive et on a la relation :

$$P_{k^2} = \Pr\left\{\chi^2 \leq k^2\right\} , \tag{6.93}$$

ce qui exprime que la probabilité cherchée est égale à la fonction de répartition de cette variable aléatoire χ^2 pour l'abscisse k^2. Afin de trouver la loi suivie par χ^2, effectuons le changement de variables aléatoires $(\boldsymbol{X} - \boldsymbol{\mu}) = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{Y}$ étudié plus haut. Il vient :

$$\chi^2 = \boldsymbol{Y}^t \boldsymbol{Y} = \sum_{i=1}^{n} Y_i^2 \,. \tag{6.94}$$

La variable aléatoire χ^2 est donc la somme des carrés de n variables aléatoires suivant la loi normale réduite. Afin de trouver l'expression analytique de la fonction de répartition F_{χ^2} de χ^2 il suffit alors de calculer l'intégrale :

$$F_{\chi^2}(u) = \Pr\left\{\chi^2 \leq u\right\} = (2\pi)^{-\frac{n}{2}} \iint\limits_{\chi^2 \leq u} \exp\{-\tfrac{1}{2}\sum_{i=1}^{n} y_i^2\}\, dy_1 \cdots dy_n \,. \tag{6.95}$$

Calculons cette intégrale en coordonnées polaires dans $\mathbb{R}^n$. Soit $r = \sqrt{\sum_i y_i^2}$ le rayon polaire et Ω l'angle solide. L'étude des intégrales multiples nous apprend d'une part que $dy_1 \cdots dy_n = r^{n-1}\, dr d\Omega$, d'où :

$$F_{\chi^2}(u) = (2\pi)^{-\frac{n}{2}} \int d\Omega \int_0^{\sqrt{u}} e^{-r^2/2} r^{n-1}\, dr\,, \tag{6.96}$$

et d'autre part, que :

$$\int d\Omega = \frac{2}{\Gamma(\frac{n}{2})} \pi^{\frac{n}{2}}\ . \tag{6.97}$$

Dans cette dernière expression, la fonction Γ est la fonction eulérienne de 2$^{\text{e}}$ espèce (voir annexe A, page 309). On obtient alors :

$$F_{\chi^2}(u) = \frac{2^{-\frac{n}{2}+1}}{\Gamma(\frac{n}{2})} \int_0^{\sqrt{u}} e^{-r^2/2}\, r^{n-1}\, dr\,. \tag{6.98}$$

Posons maintenant $t = r^2$. Il vient :

$$F_{\chi^2}(u) = \frac{1}{2\Gamma(\frac{n}{2})} \int_0^{u} e^{-\frac{t}{2}} \left(\frac{t}{2}\right)^{\frac{n}{2}-1} dt\,. \tag{6.99}$$

On trouve la densité de probabilité de la variable aléatoire χ^2 en dérivant F_{χ^2}, soit :

$$f_{\chi^2}(u) = \frac{1}{2\Gamma(\frac{n}{2})} \left(\frac{u}{2}\right)^{\frac{n}{2}-1} \exp\Big\{-\frac{u}{2}\Big\}, \qquad u \geq 0\,. \tag{6.100}$$

Une loi possédant cette densité de probabilité est dite *loi du χ^2 à n degrés de liberté.*

6.3.10 Contenu en probabilité de l'ellipsoïde d'égale densité.

A partir de la loi du χ^2 on trouve le contenu en probabilité P_{k^2} des ellipsoïdes d'égale densité $Q(\boldsymbol{x}) = k^2$, par :

$$P_{k^2} = \Pr\left\{\boldsymbol{X} | Q(\boldsymbol{X}) \leq k^2\right\} = F_{\chi^2}(k^2), \tag{6.101}$$

où F_{χ^2} est la fonction de répartition d'une loi du χ^2 dont le nombre de degrés de liberté n est égal à la dimension de l'espace où se répartit la variable aléatoire normale $\boldsymbol{X}$. Dans le cas particulier de la loi normale 3D, l'expression (6.101) prend une forme plus simple :

$$P_{k^2} = 2\Phi(k) - 1 - \sqrt{\frac{2}{\pi}}\, k \exp\{-\tfrac{1}{2}k^2\}, \tag{6.102}$$

où Φ est la fonction de répartition de la loi normale 1D réduite.

6.3.11 Introduction au test du χ^2.

Dans la pratique on utilise la formule (6.101) en sens inverse, et elle sert de base à la détection d'un signal noyé dans un bruit connu. On se donne a priori la probabilité γ pour qu'un point tiré au hasard suivant la loi normale nD, tombe dans l'ellipsoïde contenant cette probabilité γ. On a alors $P_{k_\gamma^2} = \gamma$ et l'on cherche ensuite la valeur k_γ qui correspond à cette probabilité γ. On l'obtient en inversant l'équation (6.101) :

$$k_\gamma^2 = F_{\chi^2}^{-1}(\gamma) \tag{6.103}$$

Puis, étant donné une observation, on calcule à l'aide de la formule (6.92) la valeur du χ^2 qui lui correspond, et on déclare avoir détecté un signal si $\chi^2 > k_\gamma^2$. Ce faisant, et si en fait il n'y a pas de signal, on commet une erreur dite de « *fausse-alarme* » avec une probabilité α égale à $1 - \gamma$.

Cette façon de procéder est connue sous le nom de « *test du* χ^2 », et plus spécifiquement dans le cas 1D et pour $\gamma \approx 0.997$, sous le nom de « *règle des 3 sigmas,* » car dans ce cas $k_\gamma = 3$. La table 6.4 donne des valeurs de k_γ pour différentes valeurs de n et de γ.

n	$k_{\gamma=0.50}$	$k_{\gamma\approx0.68}$	$k_{\gamma=0.90}$	$k_{\gamma=0.99}$	$k_{\gamma\approx0.997}$	$k_{\gamma=0.999}$
1	0.67449	1	1.64485	2.57583	3	3.29053
2	1.17741	1.51517	2.14597	3.03485	3.43935	3.71692
3	1.53817	1.87796	2.50028	3.36821	3.76250	4.03314
4	1.83213	2.17244	2.78916	3.64372	4.03130	4.29730
5	2.08601	2.42644	3.03914	3.88411	4.26677	4.52935
6	2.31260	2.65300	3.26261	4.10023	4.47907	4.73896
7	2.51909	2.85941	3.46656	4.29829	4.67403	4.93172
8	2.71000	3.05023	3.65535	4.48221	4.85537	5.11121
9	2.88840	3.22852	3.83193	4.65467	5.02562	5.27988
10	3.05644	3.39647	3.99840	4.81760	5.18663	5.43951
20	4.39743	4.73670	5.33029	6.12913	6.48692	6.73162
30	5.41627	5.75509	6.34476	7.13388	7.48616	7.72678

TAB. 6.4: *Table donnant le seuil k_γ au delà duquel, suivant le test du χ^2, on refuse l'hypothèse que l'observation est issue d'une loi normale à n dimensions. Le test suppose n connu et γ donné. On donne ici des valeurs de k_γ pour $\gamma =$* 0.50, 0.68, 0.90, 0.99, 0.997 *et* 0.999. *La probabilité de commettre une erreur de type* fausse-alarme *est de $1 - \gamma$. Quand $n \to \infty$, $k_\gamma^2 \to \frac{1}{2}(x_{1-\gamma} + \sqrt{2n-1})^2$, où $x_{1-\gamma}$ est un quantile de la loi normale réduite.*

6.4 Aspects numériques.

6.4.1 Quantiles de la loi normale réduite.

Le quantile Q_α de la loi normale réduite est défini par l'équation $Q_\alpha = \Phi^{-1}(1-\alpha)$, où Φ est la fonction de répartition de la loi normale réduite. Par définition, la quantité α est la probabilité pour qu'une variable aléatoire dépasse

le seuil Q_α. Le quantile d'une loi quelconque, de moyenne μ et d'écart type σ est égal à $\sigma Q_\alpha + \mu$. Ce résultat s'applique en particulier à la loi normale.

Nous prendrons comme approximation de Φ^{-1} une formule donnée par Abramowitz et Stegun (1970) [2]. Cette approximation atteint une précision absolue de $|\epsilon(\alpha)| < 4.5 \times 10^{-4}$:

$$Q_\alpha = \begin{cases} t - R(t) + \epsilon(\alpha), & t = \sqrt{-2\ln\alpha} \qquad 0 < \alpha \leq 0.5\,, \\ -t + R(t) + \epsilon(\alpha), & t = \sqrt{-2\ln(1-\alpha)} \quad 0.5 < \alpha \leq 1.0\,, \end{cases}$$

$$R(t) = \frac{c_0 + c_1 t + c_2 t^2}{1 + d_1 t + d_2 t^2 + d_3 t^3}\,,$$

$$c_0 = 2.515517, \quad c_1 = 0.802852, \quad c_2 = 0.010328\,,$$

$$d_1 = 1.432788, \quad d_2 = 0.189269, \quad d_3 = 0.001308\,.$$

6.4.2 Génération d'un couple de variables aléatoires suivant la loi normale 2D.

A partir d'un couple de variables aléatoires X_1, X_2 où chacun des termes suit la loi normale réduite, nous désirons former un couple Y_1, Y_2 suivant la loi normale 2D de moyenne $\boldsymbol{\mu}$ et de matrice des variances-covariances $\boldsymbol{V}$. En application directe du changement de variable (6.66) on trouve :

$$Y_1 = X_1\lambda_1\cos\phi - X_2\lambda_2\sin\phi + \mu_1$$
$$Y_2 = X_1\lambda_1\sin\phi + X_2\lambda_2\cos\phi + \mu_2\,,$$

où λ_1^2, λ_2^2 sont données par la formule (6.25) et $\cos\phi, \sin\phi$ par la formule (6.28) ou (6.29). On a simulé, à l'aide de ce changement de variables, les points de la figure 6.5.

6.4.3 Simulation de vecteurs suivant la loi normale nD.

Le programme `RNORMND(X,N,NP, MU,V,SEED, LAMBDA,U,FIRST)` retourne dans `X` un ensemble de `N` nombres aléatoires suivant la loi normale de moyenne `MU` et de matrice des variances-covariances `V`. Le programme retourne également dans `LAMBDA` les éléments diagonaux de la matrice $\boldsymbol{\Lambda}$, et dans `U` les vecteurs propres (composantes principales) de `V`. Les éléments de `LAMBDA` sont les valeurs singulières de `V`, c'est-à-dire les racines carrées des valeurs propres de `V`. Le paramètre `NP` est la taille physique des tableaux `LAMBDA(NP),V(NP,NP),U(NP,NP)`. La variable `SEED` initialise le tirage aléatoire. Elle doit être égale soit à un nombre premier assez grand (>1000), soit à 0 auquel cas la série de nombres aléatoires sera répétitive. Au premier appel à `RNORMND` la variable logique `FIRST` doit être égale à `.TRUE.`, on doit ensuite lui donner la valeur `.FALSE.` tant que la matrice `V` reste inchangée. Le programme `JACOBI` est extrait de *"Numerical Recipes,"* Press *et al.*(1986) [60].

```
SUBROUTINE RNORMND(X, N,NP, MU,V,SEED, LAMBDA,U,FIRST)
INTEGER*4 N, NP, SEED, I, J, NROT
LOGICAL*4 FIRST
REAL*4    X(1), MU(1), V(NP,NP), Y(1), RNORM
REAL*4    LAMBDA(1), U(NP,NP), SUM

IF (FIRST) THEN  ! Diagonalisation
```

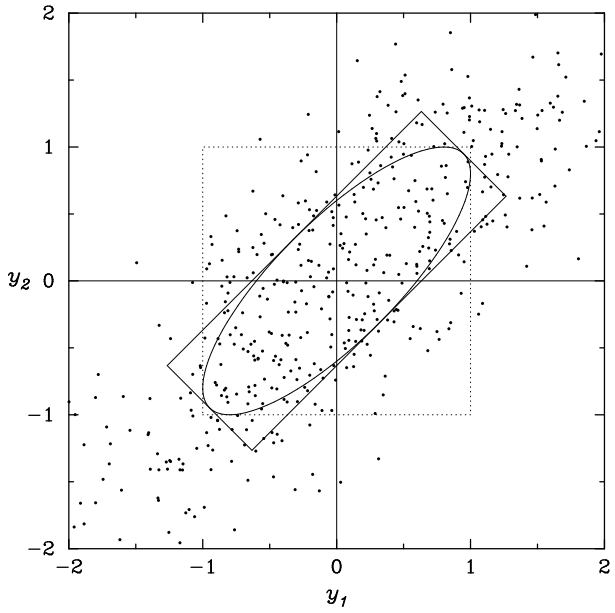


FIG. 6.5: *Simulation de points suivant la loi normale 2D de moyenne* $\boldsymbol{\mu} = 0$ *de variances* $\sigma_1^2 = \sigma_2^2 = 1$, *et de coefficient de corrélation* $\rho = 0.8$. *On a tracé l'ellipse de corrélation, le rectangle de dispersion (en trait pointillé) et le rectangle de dispersion correspondant aux axes principaux (en trait plein). On a tiré 500 points suivant cette loi. On en attendait 196.7 en moyenne dans l'ellipse de corrélation, ici il y en a 181.*

```
 CALL JACOBI(V,N,NP, LAMBDA,U,NROT)
 DO I=1,N
  LAMBDA(I) = SQRT(LAMBDA(I))
 ENDDO
ENDIF
DO I=1,N
 Y(I) = LAMBDA(I)*RNORM(SEED)
ENDDO
DO I=1,N
 SUM = 0.0
 DO J=1,N
  SUM = SUM + U(I,J)*Y(J)
 ENDDO
 X(I) = SUM + MU(I)
ENDDO
RETURN
END
```

Le programme ci-dessus a été utilisé (voir fig. 6.6) afin de simuler des vecteurs aléatoires de dimension 150, de moyenne nulle ($\boldsymbol{\mu} = 0$) et de matrice $\boldsymbol{V}$ dont les éléments v_{ij} sont tels que $v_{ij} = \sigma^2 \rho^{|i-j|}$. La matrice $\boldsymbol{V}$ a donc la structure

suivante :

$$\boldsymbol{V} = \begin{pmatrix} \sigma^2 & \sigma^2\rho & \sigma^2\rho^2 & \ddots & \sigma^2\rho^{149} \\ \sigma^2\rho & \sigma^2 & \sigma^2\rho & \ddots & \ddots \\ \sigma^2\rho^2 & \sigma^2\rho & \sigma^2 & \ddots & \ddots \\ \ddots & \ddots & \ddots & \ddots & \ddots \\ \sigma^2\rho^{149} & \ddots & \ddots & \ddots & \sigma^2 \end{pmatrix} . \tag{6.104}$$

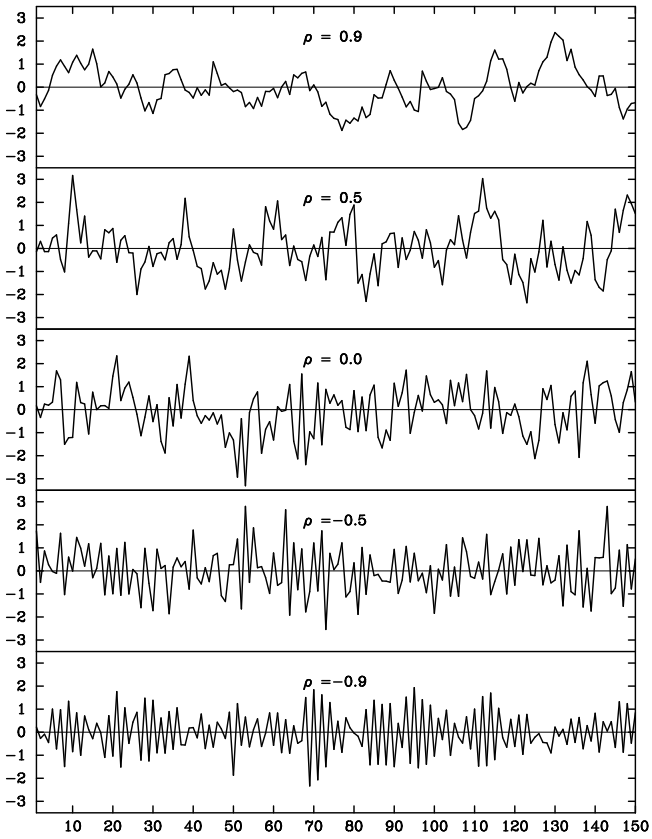


FIG. 6.6: *Simulation de vecteurs suivant la loi normale nD pour* $n = 150$. *La moyenne* $\boldsymbol{\mu}$ *est nulle et la matrice des variances-covariances* $\boldsymbol{V}$ *est donnée par l'équation* (6.104) *pour des valeurs de* $\rho = -0.9, -0.5, 0, 0.5$ *et* 0.9. *Quand* $\rho = 0$, *on a un ensemble de 150 variables aléatoires non-corrélées (et donc indépendantes dans le cas normal) ; c'est ce que l'on appelle un* « bruit blanc ».

6.5 Exercices et problèmes.

Exercice 6.1. Montrer que si les composantes d'un vecteur aléatoire $\boldsymbol{X} = (X_1, \ldots, X_n)$ suivant la loi normale à n dimensions sont 2 à 2 indépendantes, alors elles sont

mutuellement indépendantes.

Exercice 6.2. Si les variables aléatoires X_1 et X_2 sont indépendantes et suivent la loi normale réduite, montrer que les variables aléatoires Y_1 et Y_2 définies par :

$$Y_1 = \exp\{-\tfrac{1}{2}(X_1^2 + X_2^2)\},$$
$$Y_2 = \frac{1}{2\pi} \arctan \frac{X_1}{X_2},$$

sont indépendantes et suivent la loi uniforme sur $[0, 1]$. (Box & Muller, 1958, [13])

Exercice 6.3. *Rapport de deux variables aléatoires normales.* Montrez que le quotient de deux variables aléatoires normales indépendantes $\mathcal{N}(\mu_1, \sigma_1^2)/\mathcal{N}(\mu_2, \sigma_2^2)$ suit une loi dont la densité de probabilité $g(y)$ est donnée par l'expression suivante :

$$g(y) = \frac{1}{\pi} \frac{\sigma_1 \sigma_2}{\sigma_1^2 + y^2 \sigma_2^2} \Big[\exp\Big\{ -\frac{1}{2} \Big(\frac{\mu_1^2}{\sigma_1^2} + \frac{\mu_2^2}{\sigma_2^2} \Big) \Big\} + \sqrt{2\pi} z \, \Phi_0(z) \exp\Big\{ -\frac{1}{2} \frac{(\mu_1 - \mu_2 y)^2}{\sigma_1^2 + \sigma_2^2 y^2} \Big\} \Big] , \tag{6.105}$$

avec

$$\Phi_0(z) = \frac{1}{\sqrt{2\pi}} \int_0^z \exp - \frac{t^2}{2} \, dt \quad \text{et} \quad z = \frac{\mu_2 \sigma_1^2 + y \mu_1 \sigma_2^2}{\sigma_1 \sigma_2 (\sigma_1^2 + \sigma_2^2 y^2)^{\frac{1}{2}}} . \tag{6.106}$$

Montrez que cette loi ne possède pas de moyenne et par conséquent pas de variance non-plus.

Exercice 6.4. *Effet de sélection.* Une expérience est décrite par un triplet de variables aléatoires : (X_1, X_2, X_3) que l'on suppose normal. Les variables X_1 et X_2 sont indépendantes entre elles mais sont corrélées avec X_3. Les coefficients de corrélations de X_1 avec X_3 et de X_2 avec X_3 sont identiques et valent ρ. Les variances de X_1, X_2 et X_3 sont respectivement σ_1^2, σ_2^2 et σ_3^2.

Donner l'expression de la matrice des variances-covariances du triplet (X_1, X_2, X_3) et celle de la matrice des variances-covariances du couple (X_1, X_2).

Les conditions expérimentales sont telles que les variables X_1 et X_2 ne sont observables que si X_3 vaut une certaine valeur x_3. Que devient alors la matrice des variances-covariances (conditionnelle) du couple (X_1, X_2) ? Montrer, en particulier, que ces deux variables apparaissent alors anti-corrélées avec $-\rho^2$ comme coefficient de corrélation.

[Note : Si ce conditionnement est ignoré de l'expérimentateur, il pourrait déduire d'observations du couple (X_1, X_2), que ces variables sont anti-corrélées alors qu'en réalité elles sont indépendantes.]

Problème 6.5. *Symétrie circulaire.* On dit qu'un couple de variables aléatoires (X, Y) possède la symétrie circulaire si sa densité de probabilité f ne dépend que de la distance à l'origine. C'est-à-dire si :

$$f(x, y) = g(r), \quad \text{avec} \quad r = \sqrt{x^2 + y^2} .$$

Montrer que si les variables aléatoires X et Y possèdent la symétrie circulaire et qu'elles sont de plus indépendantes, alors elles suivent chacune une loi normale de moyenne nulle et de même variance (Papoulis, p.133, [53]).

Chapitre 7

Inégalités et convergences.

Ce chapitre traite de séquences de variables aléatoires $(X_1, \ldots, X_n)$ appartenant à un même espace probabilisé et dont l'indice n est indéfini, c'est-à-dire en pratique : aussi grand que l'on veut. Nous noterons simplement $\{X_n\}$ une telle suite. Nous voulons donner un sens à des expressions telles que : « la variable aléatoire X_n tend vers la variable aléatoire X », ou encore « la loi suivie par la variable aléatoire X_n tend vers une loi normale lorsque $n \to \infty$. » Pour que ces expressions (et d'autres similaires) aient un sens, il est nécessaire de préciser la notion de convergence dite *convergence stochastique* d'une variable aléatoire vers une autre.

Pour trouver les liens qui existent entre les différents types de convergences, nous avons besoin au préalable d'établir certaines inégalités.

7.1 Inégalités.

Nous n'établissons ici que des inégalités relatives à la théorie de la convergence, on trouvera au chapitre 5.2 d'autres inégalités portant sur des espérances et qui sont elles aussi d'une grande importance pratique et théorique.

7.1.1 L'inégalité de Markov.

Nous établissons d'abord le théorème de Markov valable pour des variables aléatoires *positives* possédant une moyenne $\mathrm{E}\{X\}$.

Théorème 7.1. *Soit η une variable aléatoire positive dont la moyenne $\mathrm{E}\{\eta\}$ existe et est non nulle. On a pour tout $\lambda > 1$:*

$$\Pr\{\eta \geq \lambda \,\mathrm{E}\{\eta\}\} \leq \frac{1}{\lambda} \tag{7.1}$$

Démonstration. Par hypothèse la moyenne existe. Soit μ cette moyenne. On a :

$$\mu = \int_0^\infty x \, dF(x) \neq 0 \, .$$

On établit ensuite les inégalités suivantes :

$$\mu \geq \int_{\lambda\mu}^\infty x \, dF(x) \geq \lambda\mu \int_{\lambda\mu}^\infty dF(x) = \lambda\mu \Pr\{\eta \geq \lambda\mu\},$$

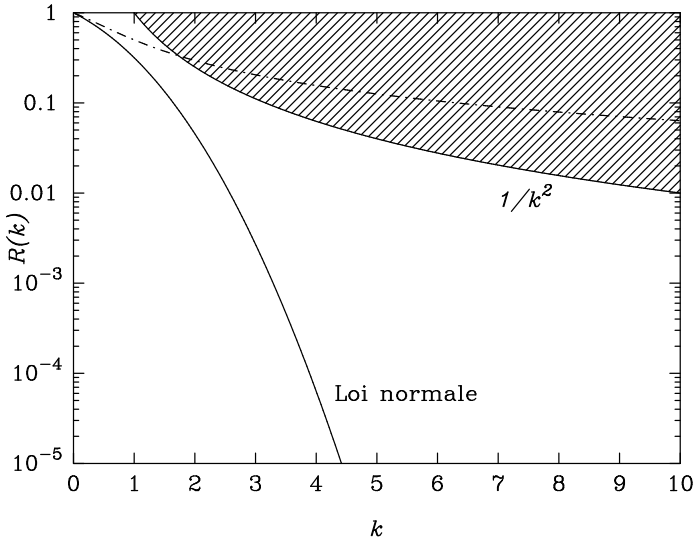


FIG. 7.1: *Graphe de la fonction d'erreur* $R(k) = \Pr\{|X-\mu|/\sigma \geq k\}$ *: probabilité résiduelle au-delà du seuil* $\mu \pm k\sigma$*, pour des lois possédant une variance* σ^2 *et donc une moyenne* μ*. D'après l'inégalité de Bienaymé-Tchébychev la fonction* $R(k)$ *est bornée supérieurement par* $1/k^2$*. On a porté, en ligne brisée, la fonction* $R(k)$ *pour la loi de Cauchy qui ne possède pas de moyenne. Pour cette loi on a posé* $\mu = 0$ *et* $\sigma = 1$*.*

ce qui démontre l'inégalité de Markov. □

7.1.2 L'inégalité de Bienaymé-Tchébychev.

Cette inégalité s'applique dès qu'une loi possède une variance. Si la variable aléatoire X possède une variance $\sigma^2 = \mathrm{Var}(X)$, elle possède aussi une moyenne $\mu = \mathrm{E}\{X\}$ et par définition son écart type est égal à $\sigma > 0$. Dans ces conditions, l'inégalité de Bienaymé-Tchébychev stipule que la probabilité pour que X s'écarte de sa moyenne de plus que k fois son écart type est inférieure à $1/k^2$. C'est-à-dire :

$$\Pr\left\{\frac{|X-\mu|}{\sigma} \geq k\right\} \leq \frac{1}{k^2}, \quad \mu = \mathrm{E}\{X\}, \sigma = [\mathrm{Var}(X)]^{\frac{1}{2}}. \tag{7.2}$$

En posant $h = k\sigma$ dans (7.2) l'inégalité s'exprime alors sous la forme suivante :

$$\Pr\{|X-\mu| \geq h\} \leq \frac{\sigma^2}{h^2}\,. \tag{7.3}$$

On démontre l'inégalité de Bienaymé-Tchébychev en posant $\eta = (X-\mu)^2$ et $\lambda = h^2$ dans l'inégalité de Markov.

Si F désigne la fonction de répartition d'une variable aléatoire X continue de moyenne μ et de variance σ^2, on a $\Pr\{|X-\mu|/\sigma \geq k\} = 1-[F(\mu+k\sigma)-F(\mu-k\sigma)]$. Introduisons la fonction d'erreur $R(k)$ égale à cette dernière expression : c'est la probabilité résiduelle au delà du seuil $\mu \pm k\sigma$. L'inégalité de Bienaymé-Tchébychev nous dit que cette probabilité est majorée par la fonction $1/k^2$.

	k	1	2	3	4
BT	$1/k^2$	1	0.25	0.1111	0.0625
BT4	$3/k^4$	(3)	0.1875	0.0370	0.01172
BT8	$105/k^8$	(105)	0.4102	0.0160	0.00160
exact	$R(k)$	0.317	0.0555	0.0027	0.000063

$R(k) = 1 - [\Phi(k) - \Phi(-k)]$

TAB. 7.1: *Bornes supérieures pour l'estimation de la fonction d'erreur* $R(k) = \Pr\{|X - \mathrm{E}\{X\}| \geq k\sigma\}$ *de la loi normale. Les bornes indiquées par BT sont fournies par l'inégalité de Bienaymé-Tchébychev aux ordres 2, 4 et 8. Lorsque la borne est supérieure à 1 (et donc inutilisable), on a porté le résultat entre parenthèses. La borne BT d'ordre* $2r$ *est donnée par* $1 \times 3 \times \ldots (2r-1)/k^{2r}$ *pour la loi normale.*

La figure 7.1 illustre cette propriété. Se placer dans le cadre des variables aléatoires quelconques (et non plus simplement continues) ne pose aucune difficulté de principe mais alourdit considérablement l'écriture à cause de la présence éventuelle de bornes discontinues pour $\mu \pm k\sigma$.

7.1.3 L'inégalité de Bienaymé-Tchébychev généralisée.

Si l'on porte dans l'inégalité de Markov $\eta = [g(X)]^r$, où g est une fonction positive, et $\lambda = h^r$, on obtient l'inégalité de Bienaymé-Tchébychev généralisée :

$$\Pr\{g(X) \geq h\} \leq \frac{\mathrm{E}\{g(X)^r\}}{h^r} \quad \text{pour} \quad g > 0 \,. \tag{7.4}$$

On retrouve (7.2) en choisissant $g(X) = |X - \mathrm{E}\{X\}|$ et $r = 2$. Pour $r = 4$ et pour des lois possédant un moment μ_4 on trouve l'inégalité de Bienaymé-Tchébychev d'ordre 4 :

$$\Pr\{|X - \mu| \geq h\} \leq \frac{\mu_4}{h^4}, \quad \mu_4 = \mathrm{E}\{(X-\mu)^4\} < \infty \,. \tag{7.5}$$

Les inégalités de Bienaymé-Tchébychev sont d'une grande importance théorique pour établir la convergence d'une suite de variables aléatoires, mais elles fournissent des bornes supérieures bien trop grandes pour être utiles dans la pratique. Ce défaut est dû à leur généralité : elles s'appliquent en effet a *toutes* les lois dès qu'elles possèdent un moment d'ordre 2.

La table 7.1 donne les bornes trouvées en utilisant ces inégalités pour la loi normale : on constate que la borne fournie est bien supérieure à la valeur exacte.

7.1.4 L'inégalité de Bernstein.

Bernstein a amélioré la borne fournie par l'inégalité de Bienaymé-Tchébychev dans le cas de variables aléatoires qui sont la somme de variables aléatoires bornées. On trouvera la démonstration de cette inégalité au chapitre 7 de l'ouvrage de Rényi [62].

Théorème 7.2. *Soient n variables aléatoires indépendantes* X_i *de moyenne* μ_i*, de variance* σ_i^2 *et bornées* $|X_i - \mu_i| \leq H$*. Soit* X *la somme des* X_i*,* $X =$

$\sum_{i=1}^n X_i$. *En tant que somme cette variable aléatoire* X *possède une moyenne* $M = \sum_{i=1}^n \mu_i$ *et une variance* $\Sigma^2 = \sum_{i=1}^n \sigma_i^2$. *Alors pour tout* $k \leq \Sigma/H$:

$$\Pr\Big\{\frac{|X-M|}{\Sigma} \geq k\Big\} \leq 2\exp\{-\frac{k^2}{2(1+kH/2\Sigma)^2}\} \tag{7.6}$$

Si les variables aléatoires indépendantes X_i *possèdent la même moyenne* μ *et la même variance* σ^2, *on a* $M = n\mu$, $\Sigma^2 = n\sigma^2$ *et il vient pour tout* $k \leq \sqrt{n}\sigma/H$:

$$\Pr\Big\{\frac{|X-n\mu|}{\sqrt{n}\sigma} \geq k\Big\} \leq 2\exp\{-\frac{k^2}{2(1+kH/2\sqrt{n}\sigma)^2}\} \tag{7.7}$$

▶**Exemple 7.1.** *Borne sur les fluctuations d'une proportion expérimentale.* Une expérience a la probabilité p de réussir et par conséquent la probabilité $q = 1-p$ d'échouer. La variable X_i indicatrice du succès vaut 1 en cas de succès et 0 en cas d'échec. On désigne par $P_n = \sum_{i=1}^n X_i/n$ la proportion des succès en n épreuves. Les variables aléatoires X_i/n sont de moyenne $\mu = p/n$, de variance $\sigma^2 = pq/n^2$ (voir en 8.1.1) et sont bornées $|X_n/n - \mu| \leq \max(p/n, q/n)$. Nous sommes alors dans les conditions d'application de la version (7.7) de l'inégalité de Bernstein où on posera $\epsilon = k\sqrt{n}\sigma$. On a alors pour $0 < \epsilon \leq \min(p,q)$:

$$\Pr\{|P_n - p| \geq \epsilon\} \leq 2\exp\{-\frac{n\epsilon^2}{2pq(1+\epsilon/2\min(p,q))^2}\}\,. \tag{7.8}$$

Cette formule limite la probabilité pour que la proportion expérimentale s'écarte de la probabilité théorique. Par exemple si $p = q = 0.5$ la probabilité pour que le nombre de succès moyen s'écarte de 0.5 de plus que $\epsilon = 0.1$ en 300 épreuves est d'après (7.8) inférieure à ≈ 0.014, ce qui s'exprime par $\Pr\{|P_{300} - 0.5| \geq 0.1\} \leq 0.014$. L'inégalité de Bienaymé-Tchébychev nous aurait fourni la borne ≈ 0.083 qui est environ 6 fois moins bonne.

7.2 La convergence stochastique.

Il existe quatre interprétations classiques de la notion de convergence stochastique : la convergence *presque-sûre*, la convergence *en moyenne quadratique*, la convergence *en probabilité* et la convergence *en loi*. Les deux premières impliquent les deux autres suivant le schéma :

$$\left.\begin{array}{r}\boxed{\text{Cv. presque-sûre}}\\ \boxed{\text{Cv. en moyenne quadratique}}\end{array}\right\} \Rightarrow \boxed{\text{Cv. en probabilité}} \Rightarrow \boxed{\text{Cv. en loi}} \tag{7.9}$$

7.2.1 La convergence en loi.

On dit qu'une variable aléatoire X_n de fonction de répartition F_n converge *en loi* vers une variable aléatoire X de fonction de répartition F, si la suite $\{F_n\}$ converge simplement vers F en tous les points où F est continue. C'est-à-dire :

$$\forall x;\, F(x^+) = F(x^-), \quad \lim_{n\to\infty} F_n(x) = F(x) \tag{7.10}$$

On notera $X_n \xrightarrow{\text{loi}} X$, si la suite X_n tend vers X suivant la convergence en loi, sous-entendu lorsque $n \to \infty$.

Cette convergence veut dire que pour tout point x où F est continue, il est possible de rendre l'erreur entre $F_n(x)$ et $F(x)$ aussi petite que l'on veut dès que n dépasse un certain rang N, soit :

$$\forall x, \forall \epsilon > 0, \exists N : [n \geq N] \Rightarrow [|F_n(x) - F(x)| \leq \epsilon] .$$

▶**Exemple 7.2.** La suite de variables aléatoires normales $X_n = N(0, \frac{\sigma^2}{n})$ converge en loi vers la variable certaine $X = 0$. La fonction de répartition de la variable certaine X est la distribution de Heaviside, elle est continue partout sauf en 0. On a bien $\forall x \neq 0$; $\lim_{n\to\infty} F_n(n) = \Phi(\sqrt{n}x/\sigma) = H(x)$ et au point de discontinuité de H on a $\forall n$, $F_n(0) = \Phi(0) = \frac{1}{2}$ alors que $H(0) = 1$.

Il faut se garder d'interprétations trop optimistes de ce type de convergence. Par exemple si la variable aléatoire X vers laquelle X_n tend *en loi* possède des moments, cela ne veut pas dire qu'à partir d'un certain n on peut approximer les moments de X_n par ceux de X. Une suite de variables aléatoires ne possédant des moments à aucun ordre peut fort bien converger en loi vers une variable aléatoire normale qui en possède à tous les ordres.

Convergence uniforme de F_n vers F.

La convergence en loi ne suppose que la convergence simple aux points de continuité de la fonction de répartition limite F, mais si F est continue, alors la convergence de F_n vers F est *uniforme*.

Théorème 7.3. *(Pólya) Si la suite $\{X_n\}$ converge en loi vers X et si la fonction de répartition F de X est continue, alors la suite $\{F_n\}$ des fonctions de répartition de X_n converge uniformément vers F.*

Cette convergence signifie que l'erreur entre $F_n(x)$ et $F(x)$ peut être rendue aussi petite que l'on veut pour *tous* les x dès que n dépasse un certain rang N, soit :

$$\forall \epsilon > 0, \exists N : [n \geq N] \Rightarrow [\forall x, |F_n(x) - F(x)| \leq \epsilon] ,$$

ou de façon équivalente :

$$\forall \epsilon > 0, \exists N : [n \geq N] \Rightarrow [\max_x |F_n(x) - F(x)| \leq \epsilon] .$$

Convergence de la fonction caractéristique.

Il est souvent plus aisé de faire appel à la fonction caractéristique afin d'établir la convergence en loi. On dispose alors des théorèmes suivants, où ce sont les théorèmes réciproques qui présentent la plus grande utilité pratique.

Théorème 7.4. *(Lévy). Si la suite de variables aléatoires $\{X_n\}$ converge en loi vers la variable aléatoire X, alors la suite des fonctions caractéristiques $\{Z_n\}$ converge* uniformément *vers la fonction caractéristique Z de X dans tout intervalle fini $[-U, U]$.*

Nous citons maintenant certains théorèmes réciproques. Dans ceux-ci, nous supposons que Z_n est la fonction caractéristique de la variable aléatoire X_n et

nous donnons quelques conditions suffisantes de convergence en loi de la suite $\{X_n\}$ vers une variable aléatoire X.

Théorème 7.5. *(Cramér, 1937). Si la suite $\{Z_n\}$ des fonctions caractéristiques converge pour tout ω vers une fonction Z continue en $\omega = 0$, alors Z est une fonction caractéristique et la suite $\{X_n\}$ converge en loi vers une variable aléatoire X possédant Z comme fonction caractéristique.*

Théorème 7.6. *(Lévy, 1922). Si la suite $\{Z_n\}$ converge* uniformément *vers Z au voisinage de $\omega = 0$, alors Z est une fonction caractéristique et la suite $\{X_n\}$ converge en loi vers une variable aléatoire X possédant Z comme fonction caractéristique.*

Théorème 7.7. *(Glivenko, 1936). Si la suite $\{Z_n\}$ converge vers une fonction Z qui est une fonction caractéristique, alors la suite $\{X_n\}$ converge en loi vers une variable aléatoire X de fonction caractéristique Z.*

Théorème 7.8. *(Dugué, 1956). Si la suite $\{Z_n\}$ des fonctions caractéristiques converge pour tout ω vers une fonction Z dont la partie réelle est continue en $\omega = 0$, alors Z est une fonction caractéristique et la suite $\{X_n\}$ converge en loi vers une variable aléatoire X possédant Z comme fonction caractéristique.*

7.2.2 La convergence en probabilité.

On dit que X_n converge vers X en *probabilité*, si :

$$\lim_{n\to\infty} \Pr\{|X_n - X| > \epsilon\} = 0. \tag{7.11}$$

On notera cette convergence : $X_n \xrightarrow{\Pr} X$. Conformément à la notion de limite, l'expression (7.11) veut dire que quel que soient $\epsilon > 0$ et $\delta > 0$ (aussi petits que l'on veut, mais non nuls), il existe un rang N au-delà duquel la probabilité pour que la variable aléatoire $X_n - X$ s'écarte de 0 à plus de ϵ peut être rendue plus petite que δ, soit :

$$\forall \epsilon > 0, \forall \delta > 0;\ \exists N \text{ tel que } [n \geq N] \Rightarrow [\Pr\{|X_n - X| > \epsilon\} < \delta] .$$

Si la variable aléatoire $X_n - X$ possède une densité de probabilité, ce type de convergence signifie que cette densité se concentre autour de l'origine lorsque $n \to \infty$: elle tend vers une distribution de Dirac.

Conditions suffisantes de convergence en probabilité.

La convergence en probabilité implique la convergence en loi (voir la démonstration dans le cours de G. Calot [14] §14), mais la réciproque n'est pas nécessairement vraie. L'exemple ci-dessous montre qu'une suite de variables aléatoires peut converger en loi vers une autre sans pour cela converger en probabilité.

►**Exemple 7.3.** La convergence en probabilité demande le calcul de la probabilité de la variable $X_n - X$, pour ce faire il faut connaître la loi du couple (X_n, X). En revanche, la convergence en loi n'exige pas cette connaissance. La suite $\{X_n\}$ peut converger en loi vers X sans même préciser si les variables X_n et X sont dépendantes

ou indépendantes. A partir de cette remarque, il est facile de construire un contre-exemple.

Considérons une suite $\{X_n\}$ et une variable aléatoire X, supposons que toutes les variables $(X, X_1, \ldots, X_n, \ldots)$ suivent la loi normale réduite et sont mutuellement indépendantes. On a alors $X_n \xrightarrow{\text{loi}} X$ (les fonctions de répartition sont les mêmes), mais $\forall n\,;\ \Pr\{|X_n - X| > \epsilon\} = 2[1 - \Phi(\frac{\epsilon}{\sqrt{2}})]$ valeur qui ne tend pas vers 0 quand $n \to \infty$. Par conséquent X_n ne tend pas vers X en probabilité.

Cependant si la suite $\{X_n\}$ converge vers une variable certaine la convergence en loi implique celle en probabilité. On a les implications suivantes, où a désigne une variable aléatoire certaine (c'est-à-dire de fonction de répartition égale à la distribution de Heaviside $H(x-a)$) :

$$[X_n \xrightarrow{\Pr} X] \Rightarrow [X_n \xrightarrow{\text{loi}} X]\,,$$
$$[X_n \xrightarrow{\Pr} a] \Leftrightarrow [X \xrightarrow{\text{loi}} a]\,.$$

Donnons d'autres conditions suffisantes de convergence en probabilité vers une variable certaine. La variable aléatoire X_n converge en probabilité vers la variable certaine a si :

- Les variables aléatoires X_n possèdent une moyenne et une variance, et cette moyenne tend vers a alors que la variance tend vers 0 :

$$[\mathrm{E}\{X_n\} \to a, \mathrm{Var}(X_n) \to 0] \Rightarrow [X_n \xrightarrow{\Pr} a]\,.$$

- Les variables X_n possèdent un moment absolu par rapport à a d'ordre supérieur à 1 et celui-ci tend vers 0 quand $n \to \infty$:

$$[\forall r \geq 1, \mathrm{E}\{|X_n - a|^r\} \to 0] \Rightarrow [X_n \xrightarrow{\Pr} a]\,.$$

  Cette condition suffisante n'est pas nécessaire : une suite $\{X_n\}$ peut fort bien converger en probabilité vers a alors qu'elle ne possède aucun moment.

7.2.3 La convergence presque-sûre.

On dit que X_n converge vers X *presque-sûrement*, si :

$$\lim_{n\to\infty} \Pr\Big\{ \bigcup_{m=0}^{\infty} |X_{n+m} - X| > \epsilon \Big\} = 0\,. \tag{7.12}$$

On notera cette convergence : $X_n \xrightarrow{\text{p.s.}} X$. Cette convergence signifie que quel que soient $\epsilon > 0$ et $\delta > 0$, il existe un N au-delà duquel la probabilité pour qu'*au moins une* des variables aléatoires $X_n - X$ s'éloigne de 0 de plus de ϵ peut-être rendue inférieure à δ. C'est-à-dire :

$$\forall \epsilon, \delta > 0, \forall M \geq 0;\ \exists N \text{ tel que } [n \geq N] \Rightarrow [\Pr\Big\{ \bigcup_{m=0}^{M} |X_{n+m} - X| \Big\} < \delta]\,.$$

L'événement $\{\bigcup_{m=0}^{\infty} |X_{n+m} - X| > \epsilon\}$ est le complémentaire de l'événement $\{\bigcap_{m=0}^{\infty} |X_{m+n} - X| \leq \epsilon\}$; si au premier est associé une probabilité $< \delta$, il est

alors associé une probabilité $\geq 1-\delta$ au deuxième. On peut donc formuler la condition de convergence presque-sûre de façon équivalente par :

$$\lim_{n\to\infty} \Pr\Big\{ \bigcap_{m=0}^{\infty} |X_{n+m} - X| \leq \epsilon \Big\} = 1 \,. \tag{7.13}$$

Cela signifie que quel que soient $\epsilon > 0$ et $\delta > 0$ il existe un N au-delà duquel la probabilité pour que *toutes* les variables aléatoires $X_n - X$ s'approchent de 0 à mieux que ϵ peut être rendue aussi proche de 1 qu'on le souhaite. C'est-à-dire :

$$\forall \epsilon, \delta > 0, \forall M \geq 0;\ \exists N \text{ tel que } [n \geq N] \Rightarrow [\Pr\Big\{ \bigcap_{m=0}^{M} |X_{n+m} - X| \leq \epsilon \Big\} \geq 1-\delta] \,.$$

Alors que la convergence en probabilité est une propriété des *lois marginales* de chaque terme de la suite $(X_n, X_{n+1}, \ldots)$ dès que l'indice n dépasse un certain N, la convergence presque-sûre est une propriété concernant la *loi conjointe* des suites $(X_n, X_{n+1}, \ldots, X_{n+M})$, pour $M \geq 0$ quelconque dès que n dépasse le rang N.

L'événement $\{\bigcap_{m=0}^{M} |X_{n+m} - X| < \epsilon\}$ est identique à l'événement $\{\max_{i\geq 0} |X_{n+i} - X| < \epsilon\}$ d'où il ressort que la convergence en probabilité de cette variable aléatoire implique, et réciproquement, la convergence presque-sûre de X_n vers X. Cette propriété s'exprime ainsi :

$$[X_n \xrightarrow{\text{p.s.}} X] \;\Leftrightarrow\; [\max_{i\geq 0} |X_{n+i} - X| \xrightarrow{\Pr} 0] \,. \tag{7.14}$$

La convergence presque-sûre implique trivialement la convergence en probabilité, mais la réciproque n'est généralement pas vraie. Une condition suffisante pour que $X_n \xrightarrow{\text{p.s.}} X$ est que la série ci-dessous converge :

$$[\sum_{n=1}^{\infty} \Pr\{|X_n - X| > \epsilon\} < \infty] \Rightarrow [X_n \xrightarrow{\text{p.s.}} X];\,. \tag{7.15}$$

7.2.4 La convergence en moyenne quadratique.

On dit que X_n converge vers X *en moyenne quadratique*, si :

$$\lim_{n\to\infty} \mathrm{E}\{(X_n - X)^2\} = 0 \tag{7.16}$$

On notera cette convergence : $X_n \xrightarrow{\text{m.q.}} X$. Le fait que la convergence en moyenne quadratique implique celle en probabilité découle de l'inégalité de Bienaymé-Tchébychev.

7.2.5 Critère de Cauchy.

Pour tous les types de convergences définis ci-dessus, il existe un critère de Cauchy correspondant. Ce critère nous permet d'établir la convergence d'une suite de variables aléatoires sans référence explicite à sa limite : il suffit de montrer qu'à partir d'un certain rang deux termes quelconques de la suite peuvent

être rendus aussi proches l'un de l'autre qu'on le souhaite. Cette condition suffisante est aussi nécessaire. Exprimons cela pour la convergence en moyenne quadratique de $\{X_n\}$ vers X. On a :

$$[\lim_{n\to\infty} \mathrm{E}\{(X_n - X)^2\} = 0] \Leftrightarrow [\forall m > 0,\ \lim_{n\to\infty} \mathrm{E}\{(X_{n+m} - X_n)^2\} = 0]\,. \quad (7.17)$$

Avec les notations introduites ci-dessus, on peut exprimer les propriétés (7.9) sous la forme suivante :

$$\left.\begin{array}{l} X_n \xrightarrow{\text{p.s.}} X \\ X_n \xrightarrow{\text{m.q.}} X \end{array}\right\} \quad \Rightarrow \quad X_n \xrightarrow{\Pr} X \quad \Rightarrow \quad X_n \xrightarrow{\text{loi}} X \qquad (7.18)$$

7.3 Lois des grands nombres.

Les lois des grands nombres établissent divers critères de convergence de la moyenne arithmétique empirique d'une suite $\{X_n\}$ vers un certain nombre. Plus précisément, la *moyenne arithmétique empirique* M_n attachée à une suite $\{X_n\}$ est une variable aléatoire ainsi définie :

$$\forall n,\ M_n = \frac{1}{n}\sum_{i=1}^{n} X_i\,. \qquad (7.19)$$

On s'intéresse alors à la convergence de $\{M_n\}$ dans quelque sens que ce soit (convergence en probabilité ou autre) vers une valeur certaine μ qui est en général l'espérance d'une certaine loi.

7.3.1 Loi des grands nombres de Bernoulli.

Cette version de la loi des grands nombre porte sur une suite d'épreuves de Bernoulli. Une suite d'épreuves $\{A_n\}$ est de Bernoulli si les épreuves sont identiques et indépendantes (comme au jeu de « pile » ou « face » par exemple).

Si p est la probabilité de succès en une épreuve, Bernoulli a montré que le nombre moyen P_n de succès en n épreuves convergeait en probabilité vers p. On a pu montrer ensuite que la convergence était presque-sûre.

Introduisons la variable aléatoire indicatrice $\mathbf{1}_{A_n}$ qui vaut 1 si l'épreuve A_n est un succès et 0 si c'est un échec. C'est une variable aléatoire dite *de Bernoulli* de moyenne p et de variance $p(1-p)$. Le nombre moyen de succès en n épreuves P_n est la moyenne arithmétique des $\mathbf{1}_{A_i}$:

$$P_n = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{A_i}\,. \qquad (7.20)$$

La loi des grands nombres de Bernoulli dit que $P_n \xrightarrow{\Pr} p$. Ce théorème est une simple conséquence de l'inégalité de Bienaymé-Tchébychev. La figure 7.2 illustre la convergence de P_n vers p.

Dans l'expression (7.20) $\sum_{i=1}^{n} \mathbf{1}_{A_i}$, le nombre de succès en n épreuves suit une loi binomiale (voir chapitre 8). Pour les variables binomiales on dispose plus généralement du théorème suivant :

Théorème 7.9. *Si les variables aléatoires $\mathcal{B}(n,p)$ suivent la loi binomiale, alors*

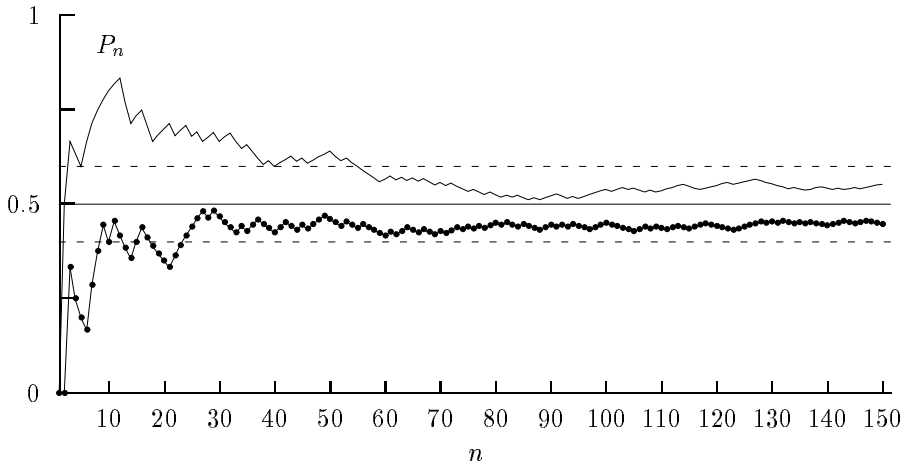


FIG. 7.2: *Illustration de la loi des grands nombres de Bernoulli. On a représenté l'évolution du nombre moyen de succès en n épreuves. La figure montre deux simulations d'une série d'épreuves identiques et indépendantes où la probabilité de succès d'une épreuve est* $p = 0.5$.

la suite $\{\frac{1}{n}\mathcal{B}(n,p)\}$ *converge presque-sûrement vers p lorsque n croît indéfiniment :*

$$[n \to \infty] \Rightarrow [\frac{1}{n}\mathcal{B}(n,p) \xrightarrow{p.s.} p]\,. \tag{7.21}$$

Ce théorème a pour conséquence que le nombre moyen de succès en n épreuves de Bernoulli converge presque-sûrement vers p : la probabilité de succès en une épreuve. Remarquons aussi que la limite p est l'espérance de $\frac{1}{n}\mathcal{B}(n,p)$.

7.3.2 Lois faibles des grands nombres.

Les lois faibles des grands nombres sont des conditions suffisantes de convergence en probabilité (convergence faible) de la moyenne arithmétique empirique M_n vers un nombre certain μ lorsque $n \to \infty$. Ces lois précisent également la nature du nombre μ. Nous commençons par une version restrictive.

Théorème 7.10. *Si les variables aléatoires* X_i *sont deux à deux indépendantes et qu'elles suivent la même loi de moyenne* μ *et de variance* σ^2*, alors leur moyenne arithmétique tend vers* μ *en probabilité :*

$$M_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{Pr} \mu\,. \tag{7.22}$$

Notons que nous n'avons supposé que l'indépendance deux à deux et non l'indépendance mutuelle. En revanche les X_i doivent suivre la *même* loi. Il existe des versions faisant appel à des hypothèses moins fortes.

Théorème 7.11. *(Markov). Si les variables aléatoires* X_i *sont deux à deux indépendantes, possèdent toutes une moyenne* μ_i *et une variance* σ_i^2 *et que de*

plus :

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \mu_i = \mu\,, \quad \textit{et} \quad \lim_{n\to\infty} \frac{1}{n}\sqrt{\sum_{i=1}^{n} \sigma_i^2} = 0\,,$$

alors M_n converge en probabilité vers μ.

Il existe une version encore plus faible où on n'exige plus l'indépendance deux à deux mais seulement la non corrélation positive forte.

Théorème 7.12. *(Bernstein) Soient des variables aléatoires X_i possédant toutes une moyenne μ_i et une variance σ_i^2. Soit ρ_{ij} le coefficient de corrélation des variables X_i et X_j (celui-ci existe nécessairement). Si les trois conditions suivantes sont satisfaites :*

1. $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \mu_k = \mu$;
2. $\forall n,\ \frac{1}{n}\sum_{i=1}^{n} \sigma_i^2 < K$ *où K est une constante indépendante de n ;*
3. $\rho_{ij} \leq R(|i-j|)$, *où $R(k)$ est une fonction non négative telle que $R(0) = 1$ et telle que* $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} R(k) = 0$;

alors la moyenne arithmétique empirique M_n converge en probabilité vers μ.

En revenant à l'hypothèse que les variables aléatoires X_i sont deux à deux indépendantes, on peut abandonner la condition d'existence des variances, il suffit que la moyenne existe.

Théorème 7.13. *(Khintchine). Si les variables aléatoires X_i sont deux à deux indépendantes et suivent la même loi de moyenne μ, alors leur moyenne arithmétique converge en probabilité vers μ.*

On peut finalement même abandonner la condition d'existence de la moyenne μ en la remplaçant par une condition d'existence de la moyenne en valeur principale.

Théorème 7.14. *(Kolmogorov). Si les variables aléatoires X_i sont deux à deux indépendantes, suivent la même loi et qu'il existe un nombre μ tel que :*

$$\mu = \lim_{L\to\infty} \int_{-L}^{L} x\, dF(x), \quad \textit{avec} \quad \lim_{x\to\infty} x(1 - [F(x) - F(-x)]) = 0\,,$$

alors la moyenne arithmétique des X_i converge en probabilité vers μ.

Ce dernier théorème veut dire que si la moyenne de la loi suivie par les X_i n'existe qu'en valeur principale et à la condition que l'erreur résiduelle tende vers 0 plus vite que $1/x$ quand $x \to \infty$, alors la moyenne arithmétique empirique tend vers cette valeur principale.

7.3.3 Lois fortes des grands nombres.

Les lois fortes des grands nombres établissent des conditions suffisantes de convergence *presque-sûre* de la moyenne arithmétique empirique vers la valeur certaine μ. Comme dans le cas des lois faibles, l'énoncé précise la nature de cette valeur qui est en général une moyenne. Nous donnons d'abord la version restrictive.

Théorème 7.15. *Si les variables aléatoires X_i sont* mutuellement *indépendantes et suivent la même loi de moyenne μ et de variance σ^2, alors leur moyenne arithmétique empirique converge presque-sûrement vers μ :*

$$M_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{p.s.} \mu\,. \tag{7.23}$$

Si les variables aléatoires X_i ne suivent pas la même loi, on a :

Théorème 7.16. *(Kolmogorov) Si les variables aléatoires X_i sont* mutuellement *indépendantes et suivent chacune une loi de moyenne μ_i et de variance σ_i^2 et si de plus la série $\sum_{k=1}^{n} \sigma_k^2/k^2$ converge, alors :*

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_i) \xrightarrow{p.s.} 0\,. \tag{7.24}$$

L'existence de la variance dans la version restrictive n'est en réalité pas nécessaire car on a :

Théorème 7.17. *(Kolmogorov) Si les variables aléatoires X_i sont* mutuellement *indépendantes et suivent la même loi de moyenne μ, alors leur moyenne arithmétique empirique converge presque-sûrement vers μ. Cette condition nécessaire est aussi suffisante et on a :*

$$[\,\forall i,\, \mathrm{E}\{X_i\} = \mu\,] \Leftrightarrow [M_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{p.s.} \mu]\,. \tag{7.25}$$

Avant de quitter les lois des grands nombres faisons remarquer que le changement de variable $\eta = X^r$ permet d'adapter ces théorèmes en des théorèmes sur la convergence des moments empiriques d'ordre r vers une valeur μ_r.

7.3.4 La loi du logarithme itéré.

Etant assuré que la suite de variables aléatoires $\{X_n\}$ satisfait aux conditions de la loi forte des grands nombres, interessons-nous aux fluctuations de M_n autour de la valeur μ vers laquelle elle converge. Plus précisément, considérons les fluctuations *réduites* de M_n, c'est-à-dire la quantité $M_n - \mu$ divisée par son écart type. Soit η_n cette variable réduite, si σ^2 est la variance des X_i, σ^2/n est la variance de M_n d'où :

$$\eta_n = \sqrt{n}\frac{M_n - \mu}{\sigma}\,.$$

L'étude de cette variable fait l'objet du chapitre § 7.4.1 suivant mais nous pouvons admettre que η_n peut prendre des valeurs arbitrairement grandes, il suffit

pour cela de choisir n assez grand. Cependant, dans le cas où les X_i sont bornés, les fluctuations extrêmes de η_n sont fortement contraintes par l'existence de *la loi du logarithme itéré*, que nous citons ci-dessous sous sa forme la plus forte.

Théorème 7.18. *(Khintchine, 1924 [42]) Si les variables aléatoires X_i sont* mutuellement *indépendantes, suivent la même loi de moyenne μ et de variance σ^2 et si, de plus, elles sont bornées, alors les valeurs extrêmes des fluctuations réduites de leur moyenne arithmétique convergent presque-sûrement vers $\pm\sqrt{2\ln\ln n}$:*

$$\min_n \sqrt{n}\frac{M_n-\mu}{\sigma} \xrightarrow{p.s.} -\sqrt{2\ln\ln n}\,, \tag{7.26a}$$

$$\max_n \sqrt{n}\frac{M_n-\mu}{\sigma} \xrightarrow{p.s.} +\sqrt{2\ln\ln n}\,. \tag{7.26b}$$

Dans ces expressions M_n désigne, comme d'habitude, la moyenne arithmétique des X_i.

▶**Exemple 7.4.** *Variable de Bernoulli.* Les issues du lancer d'une pièce de monnaie : 0 pour pile et 1 pour face, constituent une suite de variables aléatoires de Bernoulli satisfaisant les conditions du théorème 7.18. Si la probabilité d'obtenir face est p, le nombre moyen M_n de « face » obtenus en n lancers est une variable aléatoire de moyenne p et de variance $p(1-p)/n$. La loi forte des grands nombres nous dit que : $M_n \xrightarrow{\text{p.s.}} \mu$ et la loi du logarithme itéré que les fluctuations réduites de M_n sont bornées presque-sûrement par $\pm\sqrt{2\ln\ln n}$, c'est-à-dire :

$$\min_n \sqrt{n}\frac{M_n-p}{\sqrt{p(1-p)}} \xrightarrow{\text{p.s.}} -\sqrt{2\ln\ln n}\,,\ \max_n \sqrt{n}\frac{M_n-p}{\sqrt{p(1-p)}} \xrightarrow{\text{p.s.}} \sqrt{2\ln\ln n}\,. \tag{7.27}$$

On a effectué (par simulation numérique) 10^8 lancers d'une pièce de monnaie. On a calculé le nombre moyen de « face » et la suite $\{\eta_n\}$ des variable réduite de ce nombre moyen. On a finalement reporté sur la figure 7.3 la suite des valeurs extrêmes $\{\eta_{\min,n}\}$ et $\{\eta_{\max,n}\}$ des η_n. Bien que la convergence de $\eta_{\min,n}$ et $\eta_{\max,n}$ vers leur limite soit presque-sûre, la figure indique qu'elle est extrêmement lente.

7.4 Théorème central limite.

Les lois des grands nombres nous renseignent sur la convergence d'une certaine somme de variables aléatoires vers une valeur limite, mais elle ne nous disent pas *comment* on tend vers cette valeur. Il serait souhaitable de connaître la loi asymptotique vers laquelle tend cette somme. Les cas où cette loi est connue fait l'objet des divers énoncés du théorème central limite [1].

Un théorème central limite établit les conditions sous lesquelles une *somme* de variables aléatoires tend *en loi* soit vers la loi normale, soit vers la loi de Poisson, soit vers la loi certaine. Dans l'éventualité où l'on peut appliquer l'un de ces théorèmes, on peut alors approximer la fonction de répartition de la somme de ces variables aléatoires par la fonction de répartition d'une certaine variable aléatoire, par exemple normale. En pratique, on est dans les conditions

1. L'expression « théorème central limite » est la traduction mot-à-mot de l'expression allemande « Zentralen Grenzwertsatz » qui signifie : théorème établissant une limite dont l'importance est centrale (c'est-à-dire grande) en théorie des probabilités. D'après Le Cam (1986) [47] cette appelation serait due au mathématicien d'origine hongroise G. Pólya (1920) [59].

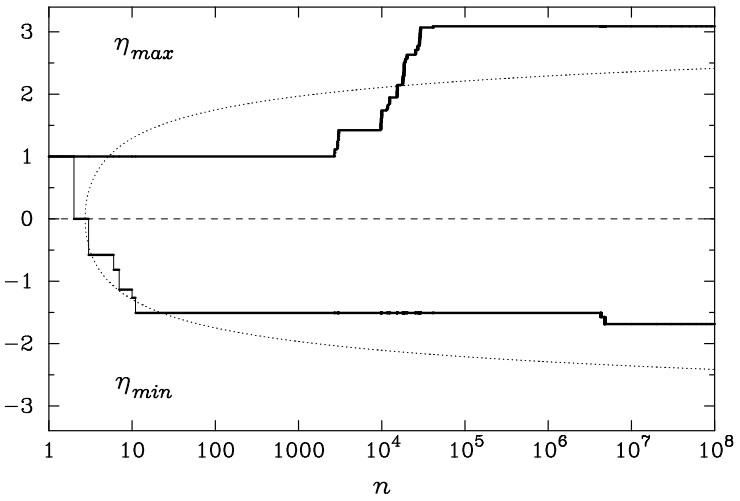


FIG. 7.3: *Illustration de la loi du logarithme itéré pour une variable aléatoire de Bernoulli (jeu de «pile» ou «face»). On a reporté le nombre d'épreuves n sur l'axe horizontal et sur l'axe vertical les valeurs extrêmes atteintes par les fluctuations réduites du nombre moyen de «face». D'après la loi du logarithme itéré, ces valeurs convergent presque-sûrement vers* $\pm\sqrt{2\ln\ln n}$. *Cette courbe limite est indiquée en pointillés sur la figure.*

d'application du théorème central limite si le nombre de variables aléatoires intervenant dans la somme croît au delà de toute limite et si chaque variable aléatoire individuelle exerce une influence arbitrairement petite au sein de cette somme.

Nous ne donnerons ici que des théorèmes limites concernant la convergence vers la loi normale. Nous noterons Φ la fonction de répartition de la loi normale réduite qui est égale, par définition, à :

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\{-\tfrac{1}{2}t^2\}\, dt\,. \tag{7.28}$$

Comme la fonction Φ est continue, d'après le théorème 7.3 de Pólya, la convergence en loi d'une suite $\{X_n\}$ vers une variable aléatoire normale réduite $\mathcal{N}(0,1)$, implique que la convergence de la fonction de répartition F_n des X_n se fait *uniformément* vers Φ.

7.4.1 Théorème central limite pour une suite de variables aléatoires indépendantes.

Afin de simplifier l'exposé, nous introduisons la notion de variable aléatoire réduite et de suite normée.

Variables réduites. Si la variable aléatoire X_i possède une moyenne μ_i et une variance σ_i^2, on appelle variable aléatoire réduite la variable aléatoire $(X_i - \mu_i)/\sigma_i$. C'est une variable aléatoire de moyenne nulle et de variance unité.

Sommes normées. Si toutes les variables X_i intervenant dans la suite $\{X_n\}$ sont indépendantes, possèdent une moyenne μ_i et une variance σ_i^2 alors nous savons que la somme $\sum_{i=1}^n X_i$ de ces variables aléatoires possède une moyenne $\sum_{i=1}^n \mu_i$, une variance $\sum_{i=1}^n \sigma_i^2$ et un écart type $\Sigma_n = [\sum_{i=1}^n \sigma_i^2]^{\frac{1}{2}}$. On fait alors correspondre à chaque terme de la suite $\{X_n\}$ une variable aléatoire η_k égale à la somme des X_i jusqu'à l'ordre k et subséquemment réduite :

$$\eta_k = \frac{1}{\Sigma_k} \sum_{i=1}^{k} (X_i - \mu_i), \quad \text{avec} \quad \Sigma_k = \sqrt{\sum_{i=1}^{k} \sigma_i^2}\,. \tag{7.29}$$

Les η_k sont les termes d'une suite $\{\eta_n\}$ appelée *suite normée* de la suite $\{X_n\}$. Les termes d'une suite normée sont de moyenne nulle et de variance unité.

Passons maintenant à l'exposé de quelques théorèmes centraux limites. C'est tout naturellement aux suites d'épreuves de Bernoulli que l'on s'est tout d'abord intéressé. Nous donnons pour commencer une version du théorème central limite les concernant.

Théorème 7.19. *(de Moivre-Laplace). Si S_n désigne le nombre de succès dans une suite d'épreuves de Bernoulli (épreuves identiques et indépendantes) chaque succès ayant la probabilité p de se réaliser, alors la variable aléatoire réduite* $\dfrac{S_n - np}{\sqrt{np(1-p)}}$ *converge en loi vers la loi normale réduite.*

Ce résultat a été établi par de Moivre (1718) et retrouvé par Laplace (1812), voir [51] et [46].

En ce qui concerne la fréquence expérimentale P_n d'apparition d'un événement dans une suite d'épreuves de Bernoulli, nous savions d'après la loi forte des grands nombres que P_n converge presque-sûrement vers la probabilité théorique p. Ce théorème nous dit maintenant que les écarts de P_n par rapport à p peuvent être approximés par une loi normale. Plus précisément on a :

$$P_n \xrightarrow{\text{p.s.}} p, \quad \text{et} \quad \sqrt{n}\frac{P_n - p}{\sqrt{p(1-p)}} \xrightarrow{\text{loi}} \mathcal{N}(0,1)\,. \tag{7.30}$$

Ce résultat nous permet d'estimer l'erreur commise en identifiant la probabilité p avec la fréquence théorique P_n. Cette estimation fait l'objet de l'exemple suivant.

▶**Exemple 7.5.** *Erreur sur une probabilité empirique.* Si l'on identifie la probabilité d'apparition d'un événement avec la fréquence expérimentale d'apparition P_n de cet événement dans une suite d'épreuves de Bernoulli, on commet alors une erreur supérieure à ϵ chaque fois que « par hasard » la suite d'épreuves nous conduit à estimer p par un nombre P_n tel que $|P_n - p| \geq \epsilon$. Pour évaluer la probabilité de cet événement malheureux il faut calculer : $\Pr\{|P_n - p| \geq \epsilon\}$. L'inégalité de Bienaymé-Tchébychev, ou mieux celle de Bernstein, nous permet de donner une borne supérieure à cette probabilité, mais le théorème central limite nous donne une bien meilleure approximation.

Introduisons la notation $q = 1 - p$, d'après (7.30) $\sqrt{n}(P_n - p)/\sqrt{pq} \xrightarrow{\text{loi}} \mathcal{N}(0,1)$. Il vient alors :

$$\Pr\{|P_n - p| \geq \epsilon\} = \Pr\Big\{\sqrt{n}\frac{|P_n - p|}{\sqrt{pq}} \geq \epsilon\sqrt{\frac{n}{pq}}\Big\}\,.$$

Si la quantité $\epsilon\sqrt{n/pq}$ est fixée et finie, c'est-à-dire si $p \neq 0$, $p \neq 1$, l'erreur ϵ et le nombre d'épreuves n sont donnés, alors le théorème de de Moivre-Laplace nous permet d'écrire que l'erreur absolue sur une fréquence expérimentale est approximativement égale à l'erreur résiduelle de la loi normale réduite au delà du seuil $\epsilon\sqrt{n/pq}$. C'est-à-dire :

$$\Pr\{|P_n - p| \geq \epsilon\} \approx 1 - [\Phi\left(\epsilon\sqrt{\frac{n}{pq}}\right) - \Phi\left(-\epsilon\sqrt{\frac{n}{pq}}\right)] = 1 - \frac{1}{\sqrt{2\pi}} \int_{-\epsilon\sqrt{n/pq}}^{\epsilon\sqrt{n/pq}} e^{-\frac{1}{2}t^2}\, dt\,. \quad (7.31)$$

Ces considérations et d'autres similaires justifient l'introduction de la notion de probabilité dans le domaine expérimental.

Nous donnons à présent une version moins restrictive du théorème central limite toujours pour des variables aléatoires identiquement réparties.

Théorème 7.20. *(Lévy-Lindeberg) Si $\{X_n\}$ est une suite de variables aléatoires* deux à deux *indépendantes, suivant toutes la même loi de moyenne μ et de variance σ^2, alors la suite normée converge en loi vers la loi normale réduite. C'est-à-dire :*

$$\frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}} \xrightarrow{loi} \mathcal{N}(0,1)\,. \quad (7.32)$$

Si les variables aléatoires intervenant dans la somme ne sont pas identiquement réparties (c'est-à-dire si elles ne suivent pas la même loi), il faut alors disposer d'un critère qui mesure l'influence d'une variable individuelle sur la somme de toutes les variables. On sera dans les conditions du théorème central limite si cette mesure tend vers zéro lorsque le nombre de termes de la somme tend vers l'infini.

Dans l'état actuel de nos connaissances nous ne disposons pas, dans le cas général, d'un tel critère qui soit une condition nécessaire et suffisante de convergence vers la loi normale ; nous ne disposons que d'une condition nécessaire : la *condition de petitesse uniforme.*

Théorème 7.21. *Si la suite $\{X_n\}$ est formée de variables aléatoires X_i indépendantes de moyennes μ_i et de variances σ_i^2, alors une condition nécessaire pour que la suite normée converge en loi vers la loi normale réduite est que :*

$$\lim_{n\to\infty} \max_{1\leq i\leq n} \frac{\sigma_i}{\Sigma_n} = 0\,. \quad (7.33)$$

Dans cette expression Σ_n est l'écart type de la somme $\sum_{i=1}^{n} X_i$.

La condition (7.33) est la condition de petitesse uniforme. Il existe des condition suffisantes et même nécessaires et suffisantes en restreignant le champ d'application du théorème central limite ; nous en donnons deux.

Théorème 7.22. *(Liapounov). Si la suite $\{X_n\}$ est formée de variables aléatoires X_i* deux à deux *indépendantes de moyennes μ_i et de variances σ_i^2 et s'il existe un nombre $\delta > 0$ tel que :*

$$\lim_{n\to\infty} \frac{1}{\Sigma_n^{2+\delta}} \sum_{i=1}^{n} \mathrm{E}\{|X_i - \mu_i|^{2+\delta}\} = 0, \quad (7.34)$$

alors la suite normée converge en loi vers la loi normale réduite. C'est-à-dire:

$$\frac{\sum_{i=1}^{n}(X_i-\mu_i)}{\left(\sum_{i=1}^{n}\sigma_i^2\right)^{\frac{1}{2}}} \xrightarrow{loi} \mathcal{N}(0,1)\,. \tag{7.35}$$

La condition (7.34) est *la condition de Liapounov*. Pour $\delta = 1$, par exemple, la condition de Liapounov s'écrit :

$$\lim_{n\to\infty} \frac{\sum_{i=1}^{n} \mathrm{E}\{|X_i-\mu_i|^3\}}{\left(\sum_{i=1}^{n}\sigma_i^2\right)^{\frac{3}{2}}} = 0 \tag{7.36}$$

La condition de petitesse uniforme (7.33) et la condition de Liapounov (7.34) sont suffisantes pour que l'on puisse approximer les écarts absolus (et donc aussi les moments), jusqu'à l'ordre $2+\delta$, de la suite normée par ceux de la loi normale réduite. C'est-à-dire :

$$\lim_{n\to\infty} \int_{-\infty}^{\infty} |x|^{2+\delta}\, dF_n^*(x) = \int_{-\infty}^{\infty} |x|^{2+\delta}\, d\Phi(x)\,, \tag{7.37}$$

où F_n^* est la fonction de répartition du terme d'ordre n de la suite normée.

Théorème 7.23. *(Lindeberg-Feller). Si la suite $\{X_n\}$ est formée de variables aléatoires X_i* deux à deux *indépendantes de moyennes μ_i et de variances σ_i^2, alors la suite normée converge en loi vers la loi normale réduite si, et seulement si, pour tout $\epsilon > 0$:*

$$\lim_{n\to\infty} \frac{1}{\Sigma_n^2} \sum_{i=1}^{n} \int_{|x-\mu_i|>\epsilon\Sigma_n} (x-\mu_i)^2 dF_i(x) = 0, \tag{7.38}$$

où F_i désigne les fonctions de répartition des variables X_i.

La condition suffisante (« si ») est due à Lindeberg, la condition (7.38) est la *condition de Lindeberg*. La condition nécessaire (« et seulement si ») est due à Feller.

7.4.2 Précision du théorème central limite.

Le théorème central limite nous permet d'établir la convergence uniforme de la suite $\{F_n^*\}$ des fonctions de répartition des variables aléatoires constituant la suite normée, mais ils ne nous dit rien quant à la qualité de l'approximation de F_n^* par Φ pour un n fini. Liapounov a examiné la vitesse de convergence de F_n^* vers Φ et ses travaux ont été améliorés par Berry et par Esseen. On dispose des deux théorèmes suivants, où la convergence vers la loi normale est assurée par les théorèmes précédents.

Théorème 7.24. *(Esseen). Soit $\{X_n\}$ une suite de variables aléatoires deux à deux indépendantes, où les X_i formant la suite possèdent une moyenne μ_i et une variance σ_i^2. S'il existe une valeur $0 < \delta \leq 1$ telle que tous les $\mathrm{E}\{|X_i-\mu_i|^{2+\delta}\}$ existent, alors on a l'inégalité :*

$$\sup_x |F_n^*(x) - \Phi(x)| \leq A \frac{1}{\Sigma_n^{2+\delta}} \sum_{i=1}^{n} \mathrm{E}\{|X_i-\mu_i|^{2+\delta}\} \tag{7.39}$$

où $\Sigma_n = (\sum_{i=1}^n \sigma_i^2)^{\frac{1}{2}}$ *est l'écart type des termes de la suite normée et* A *est une constante.*

Pour $\delta = 1$ l'inégalité (7.39) est l'inégalité d'Esseen. Si l'on pose $\delta = 1$ et si les variables aléatoires X_i sont également réparties, on obtient alors :

Théorème 7.25. *(Berry-Esseen). Si les variables aléatoires* X_i *d'une suite* $\{X_n\}$ *suivent la même loi de moyenne* μ*, de variance* σ^2 *et d'écart absolu d'ordre trois* e_3 $(\forall i,\ e_3 = \mathrm{E}\{|X_i - \mu_i|^3\})$*, alors:*

$$\sup_x |F_n^*(x) - \Phi(x)| \leq B\frac{e_3}{\sigma^3\sqrt{n}}\,, \tag{7.40}$$

où B *est une constante.*

L'inégalité (7.40) est l'inégalité de Berry-Esseen. D'après les calculs actuels on sait seulement que :

$$\frac{1}{\sqrt{2\pi}}(\approx 0.3989) \leq A \leq 0.9051 \quad \text{et} \quad \frac{1}{\sqrt{2\pi}}(\approx 0.3989) \leq B \leq 0.82\,. \tag{7.41}$$

7.5 Exemples.

7.5.1 Méthode de Monte-Carlo.

Nous allons exposer ci-dessous le principe de l'intégration par la méthode de Monte-Carlo à l'aide d'un exemple simple. Supposons que l'on désire évaluer numériquement l'intégrale suivante :

$$J = \int_0^{\frac{\pi}{2}} \cos u\, du\,, \tag{7.42}$$

dans ce cas particulier le résultat est trivial on a $J = 1$, mais le principe de la méthode s'applique à des calculs d'intégrales bien plus complexes.

En écrivant l'équation précédente sous la forme :

$$\frac{2}{\pi}J = \frac{2}{\pi}\int_0^{\frac{\pi}{2}} \cos u\, du\,, \tag{7.43}$$

on peut interpréter $\frac{2}{\pi}J$ comme l'espérance mathématique de la variable aléatoire $X = \cos(U)$ où U est une variable aléatoire uniforme entre 0 et $\frac{\pi}{2}$: $U = \mathcal{U}(0, \frac{\pi}{2})$.

Soient $(X_1, \ldots, X_n)$ n variables aléatoires calculées, grâce au changement de variable $x = \cos u$, à partir d'un ensemble $(U_1, \ldots, U_n)$ de n variables aléatoires indépendantes suivant la loi uniforme entre 0 et $\frac{\pi}{2}$. Ces variables X_i sont indépendantes, elles ont toutes la même moyenne $\mathrm{E}\{X_i\} = \frac{2}{\pi}J$ et cette moyenne existe. On définit alors les variables aléatoires J_n par l'intermédiaire de la moyenne arithmétique des X_i :

$$\frac{2}{\pi}J_n = \frac{1}{n}\sum_{i=1}^n X_i\,. \tag{7.44}$$

D'après la loi forte des grands nombres ces variables convergent presque-sûrement vers la moyenne commune des X_i lorsque $n \to \infty$:

$$\frac{2}{\pi}J_n \xrightarrow{\text{p.s.}} \mathrm{E}\{X_i\} = \frac{2}{\pi}J\,. \tag{7.45}$$

Ce résultat contient le principe même des méthodes Monte-Carlo : on approxime une certaine quantité J (ici une intégrale) par des variables aléatoires J_n à la condition que la loi des grands nombres s'applique. Pour quantifier la qualité de l'approximation pour n fixé ou pour calculer n afin que l'approximation soit meilleure qu'une certaine tolérance, il faut que les J_n possèdent une variance de façon à pouvoir utiliser le théorème central limite.

Dans notre cas les X_i possèdent la même moyenne mais aussi la même variance $\mathrm{Var}(X_i) = \sigma^2 \neq 0$. On peut alors appliquer la version de Lévy-Lindeberg du théorème central limite (théorème 7.20) qui décrit comment se répartissent les erreurs $J_n - J$ lorsque $n \to \infty$:

$$\sqrt{n}\frac{\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{2}{\pi}J}{\sigma} = \frac{2}{\pi}\frac{\sqrt{n}}{\sigma}(J_n - J) \xrightarrow{\text{loi}} \mathcal{N}(0,1)\,. \tag{7.46}$$

Pour un n assez grand on a alors quel que soit n :

$$\Pr\Big\{\frac{2\sqrt{n}}{\pi\sigma}|J_n - J| < \epsilon\Big\} \approx \Phi(\epsilon) - \Phi(-\epsilon) = 2\Phi(\epsilon) - 1\,, \tag{7.47}$$

où Φ est la fonction de répartition de la loi normale réduite. Reste à calculer σ, c'est-à-dire :

$$\sigma^2 = \frac{2}{\pi}\int_0^{\frac{\pi}{2}} (\cos x - \tfrac{2}{\pi}J)^2\, dx\,, \tag{7.48}$$

qui, dans un cas plus complexe, serait une intégrale aussi difficile à calculer que l'intégrale originale, on devrait alors se contenter d'une approximation. Ici le résultat est simple : $\sigma^2 = \frac{1}{2} - \frac{4}{\pi^2}$.

Si l'on désire une approximation de J meilleure que 5% dans 90% des cas il faut trouver n tel que : $\Pr\{|J_n - J| < 0.05\} = 0.9$. Cet objectif sera atteint dès que l'on aura :

$$\frac{\pi\sigma}{2\sqrt{n}}\epsilon < 0.05\,. \tag{7.49}$$

On a $2\Phi(\epsilon) - 1 = 0.9$ pour $\epsilon \approx 1.645$, et la condition précédente s'écrit :

$$n > \left(\frac{\pi^2}{2} - 4\right)\left(\frac{1.645}{2 \times 0.05}\right)^2 \approx 252.96\,, \tag{7.50}$$

et l'approximation souhaitée sera réalisée avec la probabilité 0.9 dès que $n > 252$.

S'il avait fallu approximer σ on aurait pu le faire à l'aide de « la formule de propagation des erreurs » (voir équation (5.67), page 78). Cette formule permet de calculer approximativement la variance de $X = \varphi(U)$ connaissant la moyenne et la variance de U :

$$\mathrm{Var}(X) \approx \mathrm{Var}(U)\left(\frac{\partial\varphi}{\partial u}\right)^2_{u=\mathrm{E}\{U\}}\,. \tag{7.51}$$

Dans l'exemple traité on a $X = \cos U$, où U est uniforme sur $[0, \frac{\pi}{2}]$, on a $\mathrm{E}\{U\} = \frac{\pi}{4}$ et $\mathrm{Var}(U) = \frac{\pi^2}{48}$ (voir équation (8.25), page 134). On en déduit $\sigma^2 \approx \frac{\pi^2}{48}\sin^2\frac{\pi}{4}$, soit $\sigma \approx 0.38$ alors que la vraie valeur est $\sigma \approx 0.31$. En utilisant cette approximation de σ on aurait trouvé que n devait être supérieur à 388 .

7.6 Exercices et problèmes.

Exercice 7.1. Soit X une variable aléatoire positive de fonction de répartition F strictement croissante. Son 3$^{\text{e}}$ quartile $x_{1/4}$ est par définition tel que $F(x_{1/4}) = 1-1/4$.

Montrer que le 3$^{\text{e}}$ quartile de X ne peut être supérieur à quatre fois son espérance mathématique.

Exercice 7.2. *Borne de Tchernov.* Par un choix approprié de la variable aléatoire η dans l'inégalité de Markov (7.1), démontrer l'inégalité de Tchernov :

$$\Pr\{X \geq \epsilon\} \leq \min_{t \geq 0} \exp[-t\epsilon + \ln \mathrm{E}\{e^{tX}\}] .$$

Si X suit une loi normale réduite calculer la borne fournie par cette inégalité lorsque ϵ vaut 3. [Rep. $e^{-9/2}$].

Exercice 7.3. Un gouvernement décide de pratiquer le contrôle des naissances de la façon suivante : chaque couple de parents a le droit d'avoir un enfant jusqu'à la naissance d'une fille ; la loi leur impose ensuite de ne plus procréer.

Quel est dans ce pays le nombre moyen de garçons et de filles sachant qu'avant ce contrôle des naissances 50% des nouveau-nés, en moyenne, étaient des garçons.

Chapitre 8

Lois de probabilité usuelles.

8.1 Lois discrètes.

8.1.1 Loi de Bernoulli.

Une variable aléatoire X est dite de Bernoulli, de paramètre p si :

$$\boxed{\Pr\{X=1\}=p\,,\qquad \Pr\{X=0\}=1-p} \tag{8.1}$$

La loi de Bernoulli sert de modèle à toute expérience dont les issues aléatoires appartiennent à deux classes mutuellement exclusives. On notera $\mathcal{B}(1,p)$ une variable aléatoire qui suit la loi de Bernoulli de paramètre p.

Fonction caractéristique.

$$Z(\omega)=(1-p)+pe^{i\omega} \tag{8.2}$$

Caractéristiques numériques de la loi de Bernoulli. Les moments non centrés μ'_k sont tous égaux à p. On a pour la moyenne et la variance :

$$\mathrm{E}\{\mathcal{B}(1,p)\}=p,\quad \mathrm{Var}(\mathcal{B}(1,p))=p(1-p). \tag{8.3}$$

8.1.2 Loi binomiale.

Une variable aléatoire X est binomiale si :

$$\boxed{\Pr\{X=k\}=C_n^k\,p^k\,(1-p)^{n-k}\qquad k=0,1,2,\ldots,n} \tag{8.4}$$

C'est une loi à deux paramètres, un paramètre entier $n>0$ et un paramètre réel $0\leq p\leq 1$; C_n^k est le coefficient du binôme. Cette loi a été introduite par Jacques Bernoulli en 1713 dans son traité « Ars Conjectandi » [7].

La loi binomiale est un modèle de n épreuves indépendantes de Bernoulli, elle donne la probabilité d'obtenir k succès en n épreuves, quand la probabilité de succès pour une épreuve est p. La variable aléatoire X représente le nombre de succès. On notera $\mathcal{B}(n,p)$ une variable aléatoire qui suit la loi binomiale de paramètres n et p.

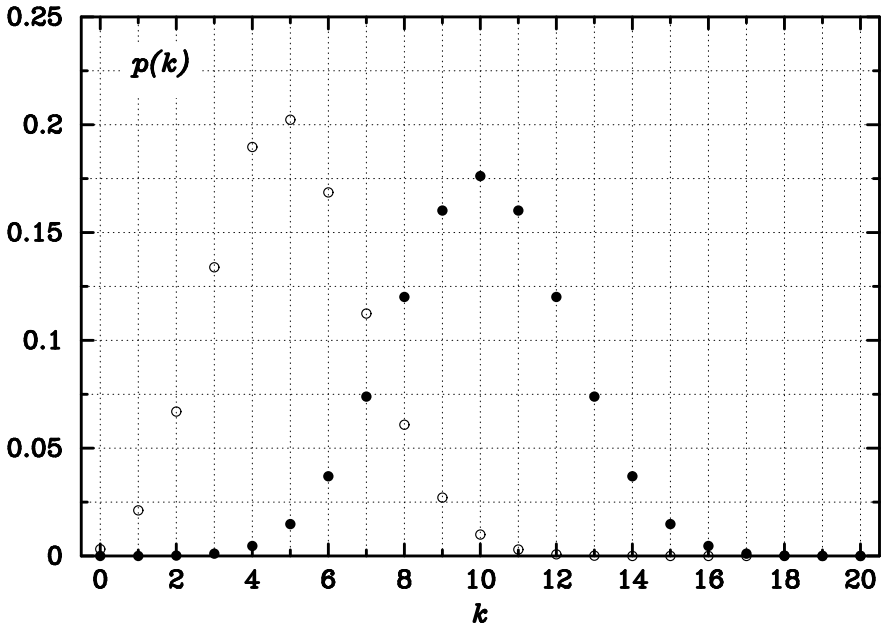


FIG. 8.1: *Répartition de la loi binomiale pour $n = 20$ et pour deux valeurs du paramètre p. Ce graphe permet, par exemple, d'évaluer la probabilité d'obtenir k fois pile en jetant $n = 20$ pièces de monnaie lorsque la probabilité pour obtenir pile est $p = 0.5$ ($\bullet$) et lorsqu'elle vaut $p = 0.25$ ($\circ$).*

Fonction de répartition. Pour tout x tel que $0 \leq x \leq n$, on a :

$$F(x) = \Pr\{X \leq x\} = \sum_{k=0}^{\lfloor x \rfloor} C_n^k p^k (1-p)^{n-k}, \tag{8.5}$$

où $\lfloor x \rfloor$ désigne la partie entière de x, c'est-à-dire le plus grand entier inférieur ou égal à x. La fonction de répartition de la loi binomiale peut aussi s'exprimer à l'aide de la fonction bêta incomplète normalisée I_p, de la façon suivante :

$$F(x) = \begin{cases} 0 & \text{si } x < 0 \\ 1 - I_p(\lfloor x \rfloor + 1, n - \lfloor x \rfloor) & \text{si } 0 \leq x \leq n \\ 1 & \text{si } x > n \end{cases} \tag{8.6}$$

Les fonctions eulériennes et les fonctions eulériennes incomplètes sont introduites dans l'appendice A.1 traitant des fonctions spéciales.

Fonction caractéristique.

$$Z(\omega) = ((1-p) + pe^{i\omega})^n \ . \tag{8.7}$$

Caractéristiques numériques de la loi binomiale.

Moments. Les moments centrés s'obtiennent grâce à la relation de récurrence :

$$\mu_0 = 1,\ \mu_1 = 0,\ \mu_{r+1} = p(1-p)(\frac{d\mu_r}{dp} + nr\mu_{r-1})\ . \tag{8.8}$$

On obtient ainsi :

$$\mu_2 = np(1-p), \quad \mu_3 = np(1-p)(1-2p)\,,$$
$$\mu_4 = np(1-p)(1-6p(1-p)+3np(1-p))\,.$$

Moyenne et variance.

$$\mathrm{E}\left\{\mathcal{B}(n,p)\right\} = np, \quad \mathrm{Var}(\mathcal{B}(n,p)) = np(1-p)\ . \tag{8.9}$$

Asymétrie et aplatissement.

$$\gamma_1 = \frac{(1-p)-p}{\sqrt{np(1-p)}}, \quad \gamma_2 = \frac{1-6p(1-p)}{np(1-p)}\ . \tag{8.10}$$

Mode. Le maximum de la loi binomiale a lieu pour la seule valeur $r = \lfloor p(n+1)\rfloor$ si $r \neq p(n+1)$, et pour les deux valeurs r et $r-1$ si $r = p(n+1)$.

Quelques propriétés.

Si les variables aléatoires X_i sont des variables aléatoires indépendantes de Bernoulli de paramètre p, la variable aléatoire $X = \sum_{i=1}^{n} X_i$ est alors binomiale de paramètres n et p, soit avec notre notation :

$$\sum_{i=1}^{n} \mathcal{B}(1,p) = \mathcal{B}(n,p)\,. \tag{8.11}$$

8.1.3 Loi de Poisson.

Une variable aléatoire X est dite de Poisson si :

$$\boxed{\Pr\{X=k\} = \frac{\mu^k}{k!}e^{-\mu} \qquad k = 0,1,2,\ldots,\infty} \tag{8.12}$$

C'est une loi à un seul paramètre réel positif $\mu > 0$, qui est un paramètre de forme. La loi de Poisson donne la probabilité de trouver exactement k événements dans un certain intervalle de temps ou à l'intérieur d'un certain domaine, quand les événements sont indépendants, arrivent à un taux constant et sont en nombre non limité. On notera $\mathcal{P}(\mu)$ une variable aléatoire qui suit la loi de Poisson de paramètre μ.

La loi qui porte son nom a été introduite par Poisson (1837) [58], elle a attendu Bortkiewicz (1898) [12] puis surtout Gosset (1907) [67] pour être appliquée. L'étude de Bortkiewicz portait sur le nombre de décès annuels par ruade de cheval dans les corps d'armée de la cavalerie prussienne. En 1910 Rutherford et Geiger [64] montrent que la désintégration α suit une loi de Poisson.

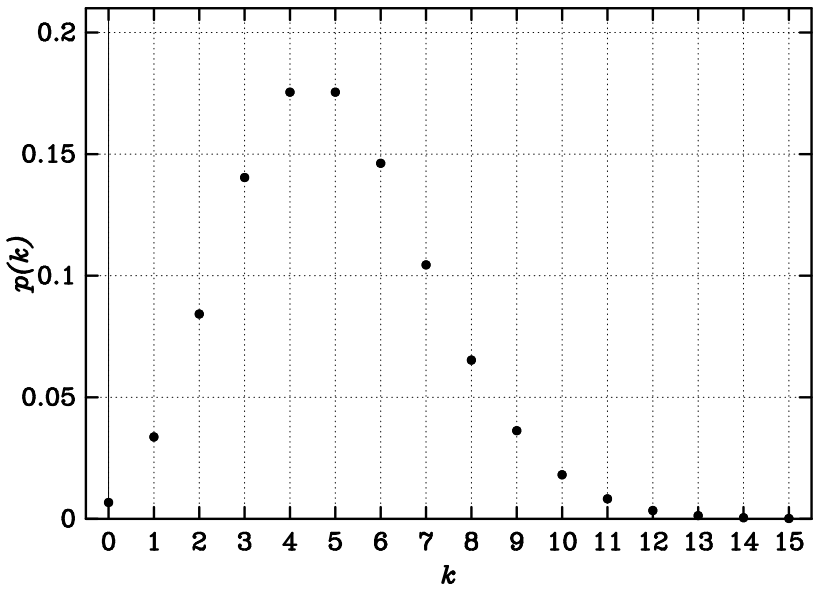


FIG. 8.2: *Répartition de la loi de Poisson de paramètre* $\mu = 5$.

Fonction de répartition. Pour tout x tel que $0 \leq x$, on a :

$$F(x) = \Pr\{X \leq x\} = \sum_{k=0}^{\lfloor x \rfloor} \frac{\mu^k}{k!} e^{-\mu}, \tag{8.13}$$

d'où :

$$F(x) = \begin{cases} 0 & \text{si } x < 0 \\ 1 - P(\lfloor x \rfloor + 1, \mu) & \text{si } 0 \leq x \end{cases} \tag{8.14}$$

Dans cette expression P représente la fonction gamma incomplète normalisée. Pour μ très grand, on a avec une bonne approximation:

$$F(x) \approx \Phi\left(\frac{x + \frac{1}{2} - \mu}{\sqrt{\mu}}\right), \tag{8.15}$$

où Φ est la fonction de répartition de la loi normale réduite.

Fonction caractéristique.

$$Z(\omega) = \exp\{\mu(e^{i\omega} - 1)\}. \tag{8.16}$$

Caractéristiques numériques de la loi de Poisson.

Moments. Les moments centrés sont donnés par la formule de récurrence :

$$\mu_0 = 1,\ \mu_1 = 0,\ \mu_{r+1} = \mu(\frac{d\mu_r}{d\mu} + r\mu_{r-1}). \tag{8.17}$$

On trouve ainsi : $\mu_2 = \mu$, $\mu_3 = \mu$, $\mu_4 = \mu + 3\mu^2$

Moyenne, variance, asymétrie et aplatissement.

$$\mathrm{E}\{\mathcal{P}(\mu)\} = \mu, \quad \mathrm{Var}(\mathcal{P}(\mu)) = \mu \quad \gamma_1 = \mu^{\frac{1}{2}}, \quad \gamma_2 = \mu^{-1}. \tag{8.18}$$

Quelques propriétés.

1. Si les variables aléatoires X_1 et X_2 sont indépendantes et suivent une loi de Poisson de paramètres respectifs μ_1 et μ_2, alors la variable aléatoire $X = X_1 + X_2$ suit une loi de Poisson de paramètre $\mu = \mu_1 + \mu_2$, soit :

$$\mathcal{P}(\mu_1) + \mathcal{P}(\mu_2) = \mathcal{P}(\mu_1 + \mu_2). \tag{8.19}$$

2. Si la moyenne $\mu = np$ d'une variable aléatoire X suivant une loi binomiale est très petite devant le nombre d'épreuves n, alors X suit approximativement une loi de Poisson de paramètre μ.

8.2 Lois continues.

8.2.1 Loi uniforme.

Une variable aléatoire X suit une loi uniforme, sur l'intervalle $[a, b[$ $(a < b)$ si elle possède une densité de probabilité $f(x)$ donnée par l'expression :

$$\boxed{f(x) = \begin{cases} \dfrac{1}{b-a} & \text{si } x \in [a, b[\\ 0 & \text{si } x \notin [a, b[\end{cases}} \tag{8.20}$$

La loi uniforme est l'analogue continu des lois discrètes décrivant des épreuves à issues équiprobables. On utilise souvent la loi uniforme avec $a = -\frac{1}{2}$ et $b = \frac{1}{2}$ pour représenter les erreurs d'arrondi dans les calculs numériques. On notera $\mathcal{U}(a, b)$ une variable aléatoire suivant la loi uniforme entre a et b.

Fonction de répartition. La fonction de répartition de la loi uniforme est donnée par la formule suivante :

$$F(x) = \begin{cases} 0 & \text{si } x \in]-\infty, a[\\ \dfrac{x-a}{b-a} & \text{si } x \in [a, b[\\ 1 & \text{si } x \in [b, \infty[. \end{cases} \tag{8.21}$$

Fonction caractéristique.

$$Z(\omega) = \frac{1}{b-a} \frac{e^{i\omega b} - e^{i\omega a}}{i\omega} \tag{8.22}$$

Caractéristiques numériques de la loi uniforme.

Moments. Les moments non-centrés μ'_r et centrés μ_r sont donnés par les formules :

$$\mu'_r = \frac{1}{r+1}\frac{b^{r+1}-a^{r+1}}{b-a} = \frac{1}{r+1}\sum_{i+j=r} a^i b^j , \tag{8.23}$$

$$\mu_{2p} = \frac{1}{2p+1}\Big(\frac{b-a}{2}\Big)^{2p},\ \mu_{2p+1} = 0 . \tag{8.24}$$

Moyenne et variance.

$$\mathrm{E}\left\{\mathcal{U}(a,b)\right\} = \frac{b+a}{2}, \quad \mathrm{Var}(\mathcal{U}(a,b)) = \frac{(b-a)^2}{12} . \tag{8.25}$$

8.2.2 Loi bêta

Une variable aléatoire X suit une loi bêta, si elle possède une densité de probabilité $f(x)$ donnée par l'expression :

$$\boxed{f(x) = \begin{cases} \dfrac{1}{B(\alpha,\beta)}\, x^{\alpha-1}(1-x)^{\beta-1} & \text{si } x \in [0,1] \\ 0 & \text{si } x \notin [0,1] \end{cases}} \tag{8.26}$$

C'est une loi à deux paramètres strictement positifs $\alpha > 0, \beta > 0$. La fonction B est la fonction eulérienne de première espèce.

La loi bêta sert de modèle aux variables aléatoires dont le domaine de définition est à support borné. Des variables aléatoires uniformes, indépendantes et triées par ordre croissant, suivent la loi bêta.

Fonction caractéristique.

$$Z(\omega) = \frac{1}{B(\alpha+\beta)}\sum_{k=0}^{\infty}\frac{(i\omega)^k}{k!}B(\alpha+k,\beta) . \tag{8.27}$$

Caractéristiques numériques.

Mode. Si $\alpha > 1, \beta > 1$ la loi bêta est unimodale et son mode (unique) vaut :

$$\frac{\alpha-1}{\alpha+\beta-2} . \tag{8.28}$$

Moments. Ils sont donnés par l'expression suivante :

$$\mathrm{E}\left\{X^k\right\} = \frac{B(\alpha+k,\beta)}{B(\alpha,\beta)} = \frac{\alpha(\alpha+1)\cdots(\alpha+k-1)}{(\alpha+\beta)(\alpha+\beta+1)\cdots(\alpha+\beta+k-1)} . \tag{8.29}$$

Moyenne et variance. De l'expression des moments on obtient :

$$\mathrm{E}\left\{X\right\} = \frac{\alpha}{\alpha+\beta}, \qquad \mathrm{Var}(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} . \tag{8.30}$$

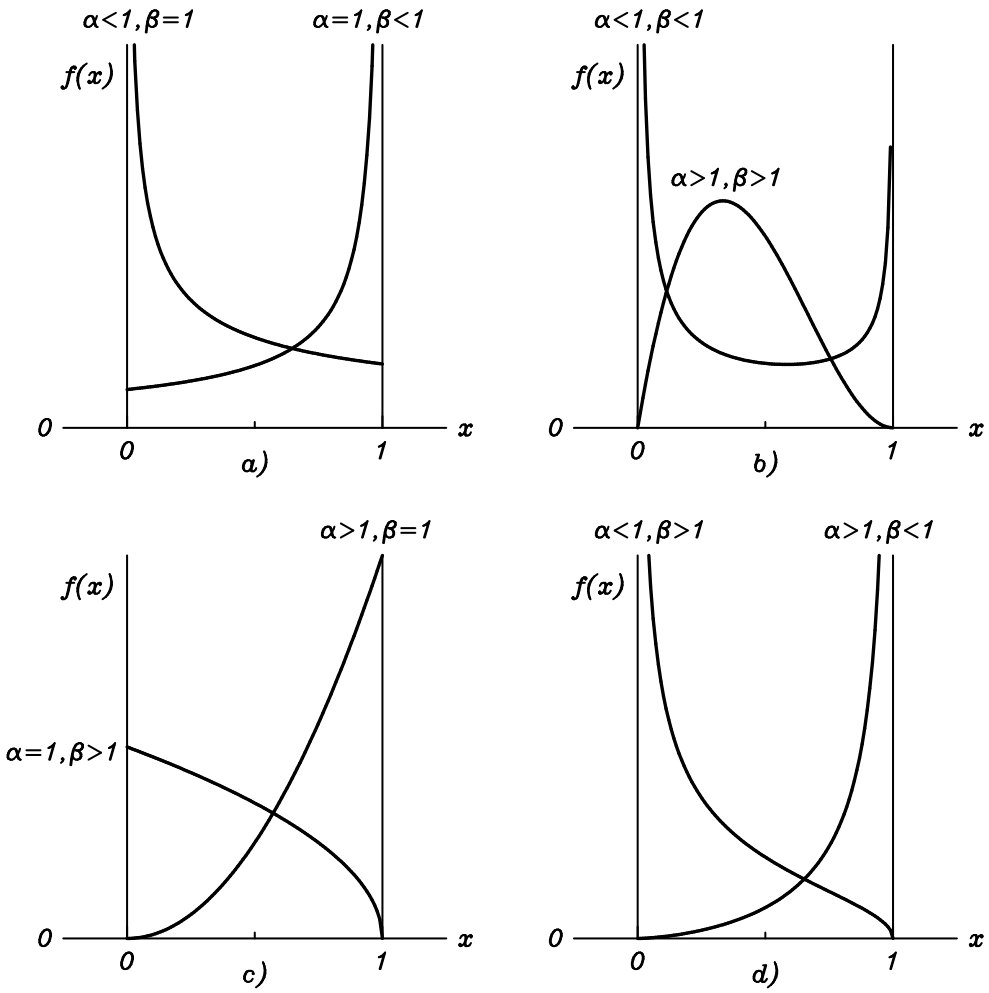


FIG. 8.3: *Diverses formes de la densité de probabilité de la loi bêta. Les graphiques correspondent aux paramètres suivants : **a)** $\alpha = 0.5$, $\beta = 1$, $\alpha = 1$, $\beta = 0.3$; **b)** $\alpha = 0.3$, $\beta = 0.5$, $\alpha = 2$, $\beta = 3$; **c)** $\alpha = 1$, $\beta = 1.5$, $\alpha = 3$, $\beta = 1$; **d)** $\alpha = 0.5$, $\beta = 1.5$, $\alpha = 2.5$, $\beta = 0.3$; pour $\alpha = 1\,\beta = 1$, la loi bêta est la loi uniforme sur* $[0, 1]$.

Quelques propriétés.

1. Si les variables aléatoires indépendantes $(X_1, \ldots, X_n)$ suivent une loi uniforme entre $[0, 1]$, et que les variables aléatoires $(X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)})$ représentent les variables X_k triées par ordre croissant, alors les variables $X_{(k)}$ suivent une loi bêta de paramètres $\alpha = k$, $\beta = n - k + 1$.

2. Si les variables aléatoires $X_i, i = 1, \ldots, m$ $Y_i, i = 1, \ldots, n$, sont indépendantes et suivent une loi normale $N(0, \sigma^2)$, alors la variable λ :

$$\lambda = \frac{\sum_{i=1}^{m} X_i^2}{\sum_{i=1}^{m} X_i^2 + \sum_{i=1}^{n} Y_i^2}, \tag{8.31}$$

suit une loi bêta de paramètres $\alpha = m/2$, $\beta = n/2$.

8.2.3 Loi du χ^2.

Une variable aléatoire X suit une loi du χ^2, si elle possède une densité de probabilité $f(x)$ donnée par l'expression :

$$\boxed{f(x) = \begin{cases} \dfrac{1}{2\Gamma(\frac{n}{2})} \left(\dfrac{x}{2}\right)^{\frac{n}{2}-1} e^{-\frac{x}{2}} & \text{si } x > 0 \\ 0 & \text{si } x \leq 0 \end{cases}} \tag{8.32}$$

C'est une loi à un seul paramètre réel n strictement positif, appelé *le degré de liberté*, qui est un paramètre de forme (voir figure 8.4). La loi du χ^2 a été introduite par l'astronome F. R. Helmert (1875 a et b) [30, 31] et le nom de « loi du χ^2 » lui a été donnée par le statisticien anglais K. Pearson. Dans le cas où $n = 3$ la loi du χ^2 est identique à la loi de Maxwell de la théorie cinétique des gaz.

Fonction de répartition. La fonction de répartition de la loi du χ^2 est donnée par l'expression suivante :

$$F(x) = \begin{cases} P(\frac{n}{2}, \frac{x}{2}) & \text{si } x \geq 0 \\ 0 & \text{si } x < 0 \end{cases} \tag{8.33}$$

où P désigne la fonction gamma incomplète normalisée (voir appendice A.1).

Fonction caractéristique.

$$Z(\omega) = (1 - 2i\omega)^{-\frac{n}{2}} . \tag{8.34}$$

Caractéristiques numériques.

Les moments centrés μ_i de la loi du χ^2 à n degrés de liberté sont donnés par la formule de récurrence suivante :

$$\mu_0 = 1, \mu_1 = n, \quad \mu_j = 2(j-1)(\mu_{j-1} + n\mu_{j-2}), \; j \geq 2 . \tag{8.35}$$

A partir des moments centrés on calcule la moyenne, la variance, l'asymétrie et l'aplatissement :

$$\mathrm{E}\{X\} = n, \quad \mathrm{Var}(X) = 2n, \quad \gamma_1 = 2\sqrt{\frac{2}{n}}, \quad \gamma_2 = \frac{12}{n}\,. \tag{8.36}$$

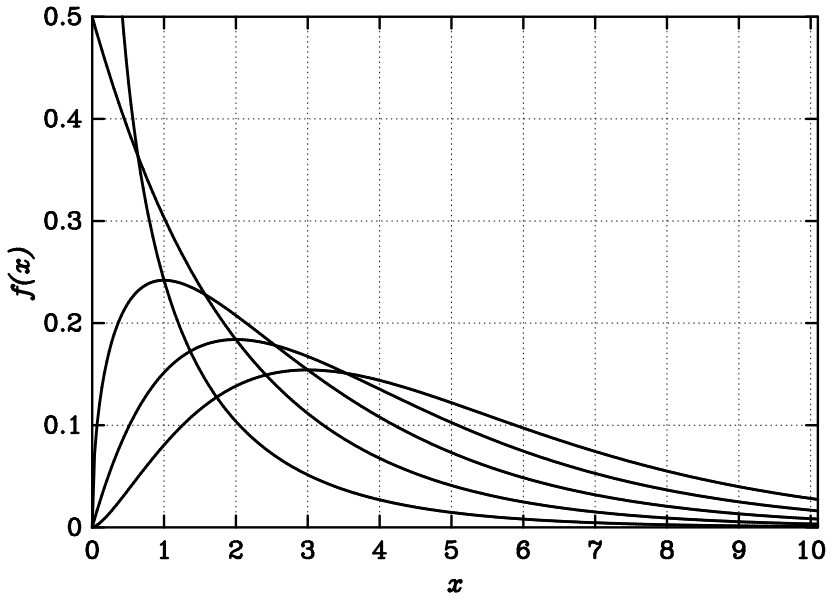


FIG. 8.4: *Densité de probabilité de la loi du χ^2 pour $n = 1, 2, 3, 4$ et 5 degrés de liberté.*

Quelques propriétés.

1. Somme des carrés de variables aléatoires normales réduites.
   Si les variables aléatoires X_i sont indépendantes et suivent la loi normale réduite, alors la variable aléatoire X^2 telle que :

   $$X^2 = \sum_{i=1}^{n} X_i^2\,,$$

   suit une loi du χ^2 à n degrés de liberté.

2. Somme de variables aléatoires.
   Il découle de la propriété précédente, que si les variables aléatoires indépendantes X_n^2 et X_m^2 suivent des lois du χ^2 à n et m degrés de liberté, alors la variable aléatoire :

   $$X_{n+m}^2 = X_n^2 + X_m^2\,,$$

   suit une loi du χ^2 à $n + m$ degrés de liberté.

3. Formule asymptotique.
   La variable aléatoire $Z_n = (2X_n^2)^{1/2} - (2n-1)^{1/2}$ tend rapidement vers la loi normale réduite quand n tend vers l'infini.

8.2.4 Loi t de Student.

Une variable aléatoire X admet une loi de Student, si elle possède une densité de probabilité $f(x)$ donnée par l'expression :

$$\boxed{f(x) = \frac{1}{\sqrt{n}B(\frac{1}{2}, \frac{n}{2})} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}} \tag{8.37}$$

C'est une loi à un seul paramètre entier $n > 0$, appelé *le degré de liberté*, qui est un paramètre de forme (voir figure 8.5). Cette loi a été introduite par le statisticien anglais W.S. Gosset en 1908 [68]. "Student" est le pseudonyme sous lequel son employeur l'avait autorisé à publier ses travaux.

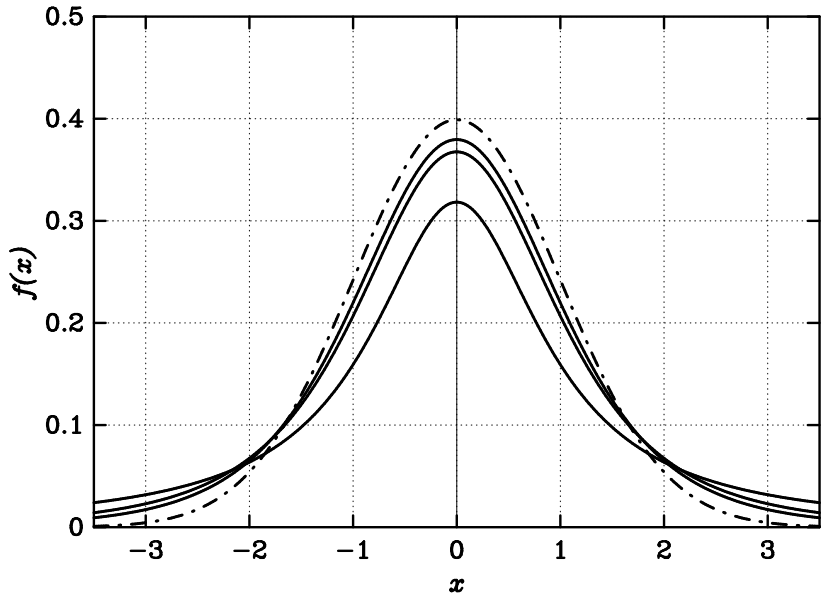


FIG. 8.5: *Densité de probabilité de la loi de Student à $n = 1, 2$ et 5 degrés de liberté. La courbe en trait interrompu est la loi normale réduite, forme asymptotique de la loi de Student pour $n = \infty$.*

Fonction de répartition. La fonction de répartition de la loi de Student est donnée par l'expression suivante :

$$F(x) = 1 - \tfrac{1}{2} I_{\frac{n}{n+x^2}}(\tfrac{n}{2}, \tfrac{1}{2}), \tag{8.38}$$

où I désigne la fonction bêta incomplète normalisée.

Caractéristiques numériques de la loi de Student.

Moments. Les moments centrés de la loi de Student sont donnés par l'expression suivante :

$$\mu_{2r} = \frac{n^r \Gamma(r+\frac{1}{2})\Gamma(\frac{n}{2}-r)}{\Gamma(\frac{1}{2})\Gamma(\frac{n}{2})}, \quad \mu_{2r+1} = 0 \quad 2r < n \ . \tag{8.39}$$

Moyenne et variance.

$$\mathrm{E}\{X\} = 0, \ n > 1; \quad \mathrm{Var}(X) = \frac{n}{n-2}, \ n > 2 \ . \tag{8.40}$$

Asymétrie et aplatissement.

$$\gamma_1 = 0, \ n > 3; \quad \gamma_2 = \frac{6}{n-4}, \ n > 4 \ . \tag{8.41}$$

Quelques propriétés.

Si X suit une loi normale réduite et Y une loi du χ^2 à n degrés de liberté et si ces variables aléatoires sont indépendantes, alors la variable $T = X/\sqrt{Y/n}$ suit une loi de Student à n degrés de liberté.

8.2.5 Loi F de Fisher.

Une variable aléatoire X admet une loi de Fisher si elle possède une densité de probabilité $f(x)$ donnée par l'expression :

$$\boxed{f(x) = \begin{cases} \dfrac{1}{B\left(\frac{\nu_1}{2}, \frac{\nu_2}{2}\right)} \left(\dfrac{\nu_1}{\nu_2}\right)^{\frac{1}{2}\nu_1} x^{\frac{1}{2}\nu_1 - 1} \left(1 + \dfrac{\nu_1}{\nu_2} x\right)^{-\frac{1}{2}(\nu_1+\nu_2)} & \text{si } x > 0 \\ 0 & \text{si } x \leq 0 \end{cases}} \tag{8.42}$$

C'est une loi à deux paramètres réels ν_1 et ν_2 strictement positifs, appelés *degré de liberté du numérateur* et *degré de liberté du dénominateur*. Ce sont des paramètres de forme (voir figure 8.6). La loi de Fisher a été introduite par Fisher en 1925 [23] elle est aussi appelée loi de Snedecor.

Fonction de répartition. La fonction de répartition de la loi de Fisher est donnée par l'expression suivante :

$$F(x) = \begin{cases} 1 - I_{\frac{\nu_2}{\nu_2+\nu_1 x}}\left(\frac{\nu_2}{2}, \frac{\nu_1}{2}\right) & \text{si } x \geq 0 \\ 0 & \text{si } x < 0 \end{cases} \tag{8.43}$$

où I désigne la fonction bêta incomplète normalisée.

Caractéristiques numériques de la loi de Fisher.

Moments non-centrés. Ils sont donnés par l'expression suivante :

$$\mu'_r = \left(\frac{\nu_2}{\nu_1}\right)^r \frac{\Gamma(\frac{1}{2}\nu_1 + r)\Gamma(\frac{1}{2}\nu_2 - r)}{\Gamma(\frac{1}{2}\nu_1)\Gamma(\frac{1}{2}\nu_2)}, \quad \nu_2 > 2r \ . \tag{8.44}$$

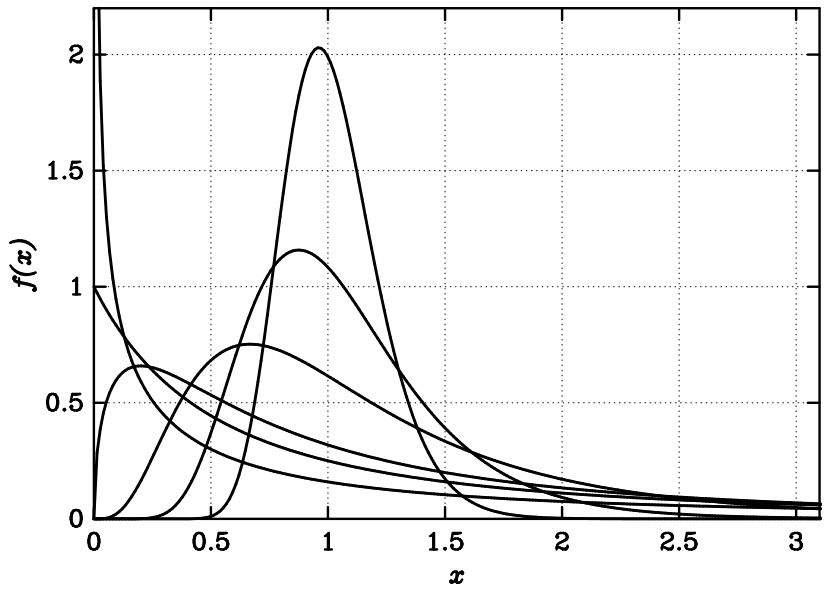


FIG. 8.6: *Densité de probabilité de la loi de Fisher pour* $\nu_1 = \nu_2 = 1, 2, 3, 10, 30$ *et* 100.

Moyenne et variance.

$$\mathrm{E}\{X\} = \frac{\nu_2}{\nu_2 - 2},\ \nu_2 > 2; \quad \mathrm{Var}(X) = \frac{2\nu_2^2(\nu_1 + \nu_2 - 2)}{\nu_1(\nu_2 - 2)^2(\nu_2 - 4)},\ \nu_2 > 4\ . \tag{8.45}$$

Quelques propriétés.

1. Si les variables aléatoires X_1 et X_2 sont indépendantes et suivent une loi du χ^2 respectivement à ν_1 et ν_2 degrés de liberté, la variable aléatoire $X = \dfrac{X_1/\nu_1}{X_2/\nu_2}$ suit une loi de Fisher à ν_1 et ν_2 degrés de liberté.

2. Si la variable aléatoire X suit une loi bêta de paramètres α, β, alors la variable aléatoire $\dfrac{\beta}{\alpha}\dfrac{X}{1-X}$ suit une loi de Fisher de paramètres $2\alpha, 2\beta$.

8.2.6 Loi exponentielle.

Une variable aléatoire X suit une loi exponentielle si elle possède une densité de probabilité $f(x)$ donnée par l'expression :

$$\boxed{f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{si } x > 0 \\ 0 & \text{si } x < 0 \end{cases}} \tag{8.46}$$

C'est une loi à un seul paramètre réel λ ($\lambda > 0$), qui est un paramètre d'échelle (voir figure 8.7).

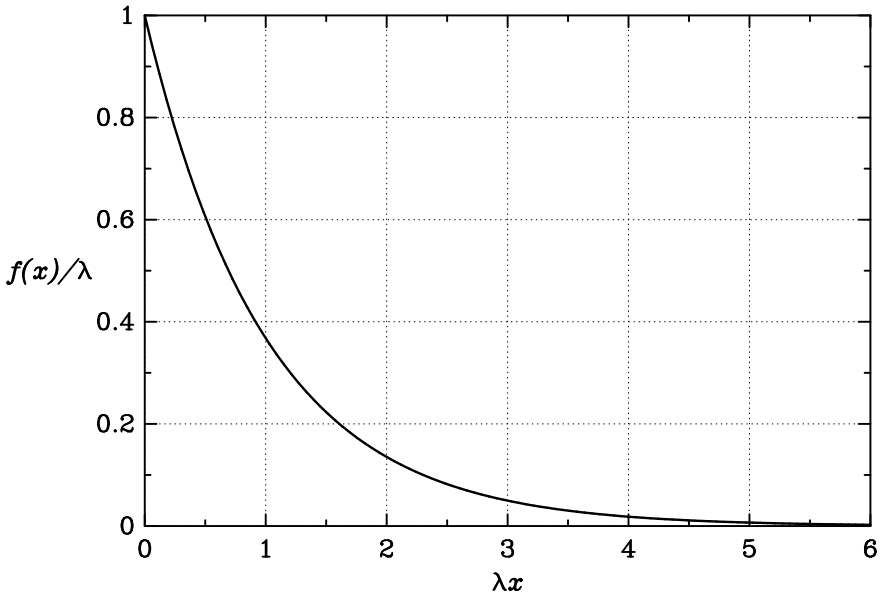


FIG. 8.7: *Densité de probabilité de la loi exponentielle.*

Fonction de répartition.

$$F(x) = \begin{cases} 1 - e^{-\lambda x} & \text{si } x \geq 0 \\ 0 & \text{si } x < 0 \end{cases} \tag{8.47}$$

Fonction caractéristique.

$$Z(\omega) = (1 - i\omega/\lambda)^{-1} \tag{8.48}$$

Caractéristiques numériques de la loi exponentielle.

Moments non-centrés. Ils sont donnés par la formule suivante :

$$\mu'_k = \frac{k!}{\lambda^k}\,, \tag{8.49}$$

Moyenne et variance.

$$\mathrm{E}\,\{X\} = \frac{1}{\lambda}\,, \quad \mathrm{Var}(X) = \frac{1}{\lambda^2}\,, \tag{8.50}$$

Asymétrie et aplatissement.

$$\gamma_1 = 2\,, \quad \gamma_2 = 6\,. \tag{8.51}$$

Mode. A proprement parler, la densité de probabilité de la loi exponentielle n'étant pas définie en 0, elle ne possède pas de mode. C'est cependant au voisinage de 0 que sa densité est maximum.

Quelques propriétés.

1. Absence de mémoire de la loi exponentielle.
   La loi exponentielle est la seule loi qui n'a *pas de mémoire*, c'est-à-dire qu'elle possède la propriété suivante

$$[\Pr\{X \leq x\} = 1 - e^{-\lambda x}] \Leftrightarrow [\Pr\{X > x + y | X > y\} = \Pr\{X > x\}]. \tag{8.52}$$

   Le terme de droite de cette équivalence est la condition d'absence de mémoire. On vérifie facilement que cette condition est nécessaire pour que la loi soit exponentielle (partie $\Rightarrow$ de l'équivalence). Le fait que c'est aussi une condition suffisante (partie $\Leftarrow$) est plus délicat et sa démonstration fait l'objet du chapitre 9.1.1.

2. Loi exponentielle et loi de Poisson.
   Soit un événement arrivant aléatoirement dans le temps avec un taux moyen λ par unité de temps. On suppose que cet événement suit une loi de Poisson, c'est-à-dire que dans un temps fini t, la probabilité d'obtenir k occurences de cet événement est donnée par la formule :

$$\Pr\{N = k\} = p_k(t) = \frac{(\lambda t)^k}{k!} e^{-\lambda t}.$$

   La probabilité de ne pas obtenir d'événement pendant le temps t est donc :

$$\Pr\{N = 0\} = p_0(t) = e^{-\lambda t}.$$

   En prenant comme variable aléatoire le temps d'arrivée de l'événement Poissonien, on obtient directement sa fonction de répartition $F(t)$:

$$\begin{aligned} F(t) &= \Pr\{X \leq t\} = 1 - \Pr\{X > t\} = 1 - p_0(t) \\ &= 1 - e^{-\lambda t} \end{aligned}$$

   et sa densité de probabilité :

$$f(t) = \frac{dF}{dt} = \lambda e^{-\lambda t}.$$

   Le temps d'arrivée d'un événement Poissonien suit donc une loi exponentielle. Nous verrons plus loin (chapitre 9) que la réciproque est également vraie.

8.2.7 Loi gamma ou loi d'Erlang.

Une variable aléatoire X admet une loi gamma, si elle possède une densité de probabilité $f(x)$ donnée par l'expression :

$$\boxed{f(x) = \begin{cases} \dfrac{\lambda^\nu}{\Gamma(\nu)} x^{\nu-1} e^{-\lambda x} & \text{si } x > 0 \\ 0 & \text{si } x \leq 0 \end{cases}} \tag{8.53}$$

C'est une loi à deux paramètres réels positifs : ν et $\lambda > 0$, ν est un paramètre de forme et λ un paramètre d'échelle (voir figure 8.8). Quand ν est un entier la fonction eulérienne Γ est égale à $(n-1)!$.

Sous des hypothèses assez générales (voir chapitre 9), le temps T qu'il faut attendre avant l'arrivée de ν photons, quand on reçoit λ photons par unité de temps, suit une loi gamma de paramètres ν et λ.

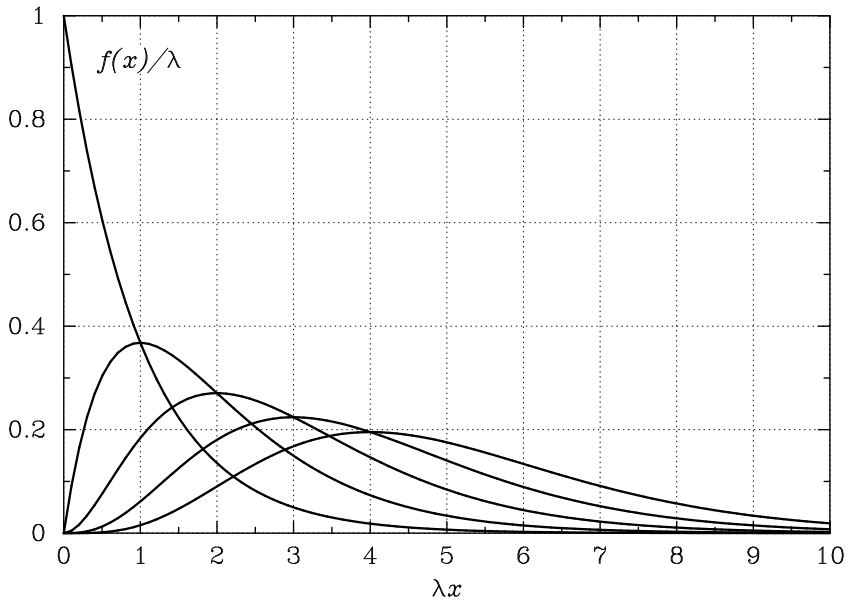


FIG. 8.8: *Densité de probabilité de la loi gamma pour le paramètre $\nu = 1, 2, 3, 4$ et 5. Pour $\nu = 1$ la loi gamma est identique à la loi exponentielle.*

Fonction de répartition. La fonction de répartition de la loi gamma est donnée par l'expression :

$$F(x) = \begin{cases} P(\nu, \lambda x) & \text{si } x \geq 0 \\ 0 & \text{si } x < 0 \end{cases} \tag{8.54}$$

où P désigne la fonction gamma incomplète normalisée.

Fonction caractéristique.

$$Z(\omega) = (1 - i\omega/\lambda)^{-\nu} \tag{8.55}$$

Caractéristiques numériques de la loi gamma.

Moments. Les moments non-centrés μ'_r existent et valent :

$$\mu'_r = \frac{\Gamma(\nu + r)}{\Gamma(\nu)} \frac{1}{\lambda^r} \, . \tag{8.56}$$

Moyenne, variance, asymétrie et aplatissement.

$$\mathrm{E}\{X\} = \frac{\nu}{\lambda}, \quad \mathrm{Var}(X) = \frac{\nu}{\lambda^2}, \quad \gamma_1 = \frac{2}{\sqrt{\nu}}, \quad \gamma_2 = \frac{6}{\nu}\,. \tag{8.57}$$

Mode. Le mode de la loi gamma est égal à $\dfrac{\nu-1}{\lambda}$ si $\nu \geq 1$ et à 0 si $0 < \nu < 1$.

Quelques propriétés.

1. Si X suit une loi gamma de paramètres ν et λ, la variable aléatoire $Y = 2\lambda X$ suit une loi du χ^2 à $n = 2\nu$ degrés de liberté.
2. Si les variables aléatoires indépendantes $X_1, \ldots, X_n$ suivent une loi exponentielle de paramètre λ, la variable aléatoire $Y = \sum_{i=1}^{n} X_i$ suit alors une loi gamma de paramètre λ et de paramètre $\nu = n$.

8.2.8 Loi log-normale.

Une variable aléatoire X admet une loi log-normale (ou logarithmiquement normale), si elle possède une densité de probabilité $f(x)$ donnée par l'expression :

$$\boxed{f(x) = \begin{cases} \dfrac{1}{\sqrt{2\pi}\sigma x} \exp\left\{-\dfrac{1}{2}\left(\dfrac{\ln x - \mu}{\sigma}\right)^2\right\} & \text{si } x > 0 \\ 0 & \text{si } x \leq 0 \end{cases}} \tag{8.58}$$

C'est une loi à deux paramètres : μ et $\sigma > 0$ (voir figure 8.9).

Fonction de répartition. La fonction de répartition de la loi log-normale est donnée par l'expression :

$$F(x) = \begin{cases} \Phi\Big(\dfrac{\ln x - \mu}{\sigma}\Big), & \text{si } x \geq 0\,; \\ 0 \qquad\quad , & \text{si } x < 0\,; \end{cases} \tag{8.59}$$

où Φ désigne la fonction de répartition de la loi normale réduite.

Caractéristiques numériques de la loi log-normale.

Moments. La loi log-normale possède des moments à tous les ordres. Les moments non-centrés sont donnés par la formule :

$$\mu'_k = \exp\left(k\mu + \tfrac{1}{2}k^2\sigma^2\right)\,. \tag{8.60}$$

Moyenne et variance. A l'aide de la formule précédente on trouve :

$$\mathrm{E}\{X\} = \exp\left(\mu + \tfrac{1}{2}\sigma^2\right)\,, \quad \mathrm{Var}(X) = (\exp\sigma^2 - 1)\exp(2\mu + \sigma^2)\,. \tag{8.61}$$

Asymétrie et aplatissement.

$$\gamma_1 = (\exp\sigma^2 - 1)(\exp\sigma^2 + 2)^2\,,$$
$$\gamma_2 = (\exp\sigma^2 - 1)(\exp 3\sigma^2 + 3\exp 2\sigma^2 + 6\exp\sigma^2 + 6)\,.$$

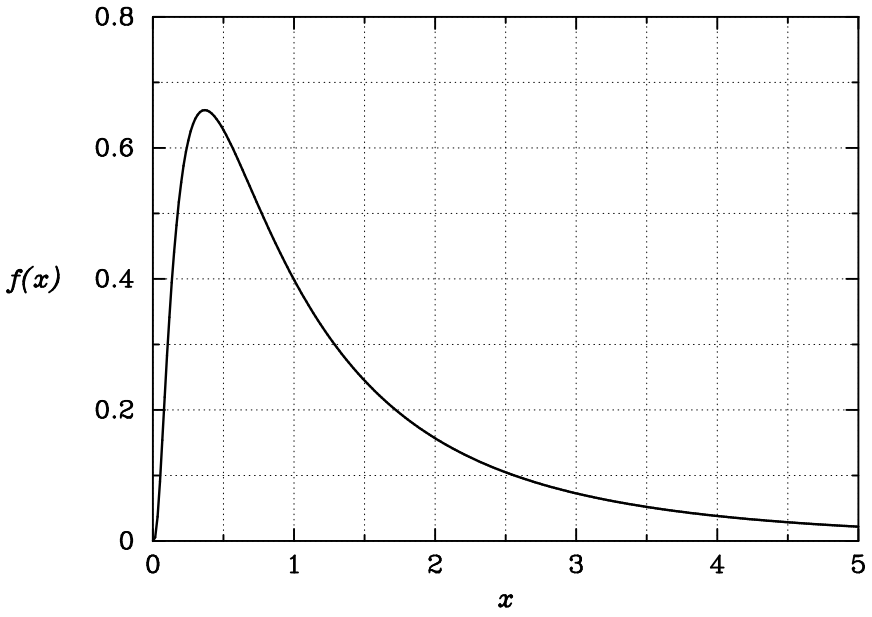


FIG. 8.9: *Densité de probabilité de la loi log-normale pour les paramètres* $\mu = 0$ *et* $\sigma = 1$.

Mode. La loi log-normale est unimodale, de mode $\exp(\mu - \sigma^2)$.

Médiane. La médiane de la loi est égale à $\exp \mu$.

Quelques propriétés.

1. D'après les valeurs précédentes, on trouve dans cet ordre : mode $<$ médiane $<$ moyenne.

2. Si la variable aléatoire X suit la loi normale réduite, la variable aléatoire X :

$$X = \exp(\sigma X + \mu) \tag{8.62}$$

suit une loi log-normale de paramètres μ et σ.

3. Si les variables aléatoires X_1 et X_2 sont indépendantes et suivent respectivement une loi log-normale de paramètres μ_1, σ_1 et μ_2, σ_2, alors le produit $X_1 X_2$ est une variable aléatoire suivant une loi log-normale de paramètres $\mu_1 + \mu_2$, $\sqrt{\sigma_1^2 + \sigma_2^2}$ et le quotient X_2/X_1 est une variable aléatoire qui suit également une loi log-normale de paramètres $\mu_2 - \mu_1$, $\sqrt{\sigma_1^2 + \sigma_2^2}$.

4. Si la variable aléatoire Y résulte de l'effet multiplicatif de n variables aléatoires X_i indépendantes strictement positives, de façon que : $Y = \prod_{i=1}^{n} X_i$, et si les X_i suivent une loi identique, il résulte du théorème central limite que Y converge en loi vers une loi log-normale. Si les X_i ne sont pas

identiquement distribués mais s'ils satisfont à la condition de petitesse uniforme (voir théorème 7.21), c'est-à-dire si la variation des $\ln X_i$ est faible devant celle de leur somme, alors Y converge également en loi vers une loi log-normale.

8.2.9 Loi de Cauchy.

Une variable aléatoire X admet une loi de Cauchy, si elle possède une densité de probabilité $f(x)$ donnée par l'expression :

$$\boxed{f(x) = \frac{1}{\pi} \frac{\lambda}{\lambda^2 + (x-\alpha)^2}} \tag{8.63}$$

C'est une loi à deux paramètres: α et λ ($\lambda > 0$), où α est un paramètre de position et λ un paramètre d'échelle (voir figure 8.10).

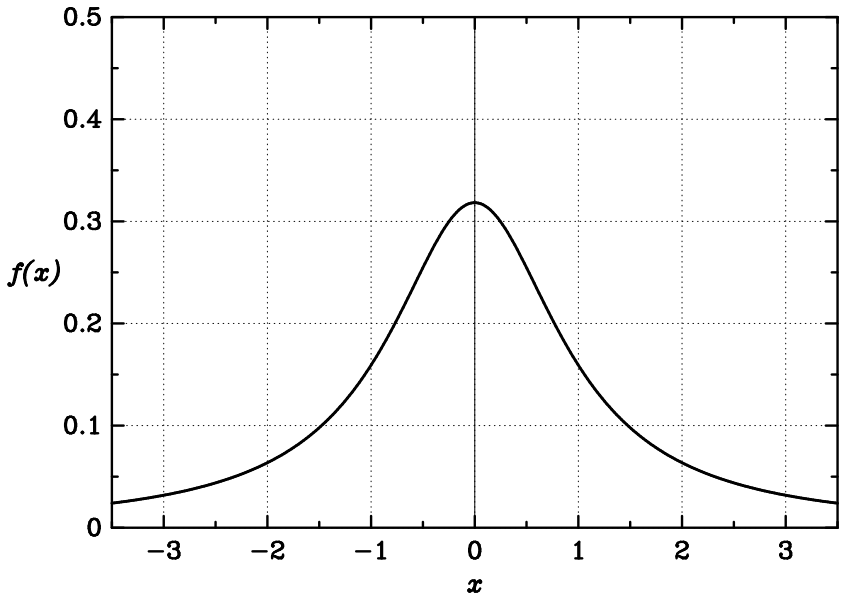


FIG. 8.10: *Densité de probabilité de la loi de Cauchy réduite, de paramètres* $\alpha = 0$ *et* $\lambda = 1$.

Cette loi qui ne possède aucune bonne propriété est utilisée pour simuler l'apparition de mesures aberrantes ou pour tester la fiabilité de certains algorithmes.

Fonction de répartition.

$$F(x) = \frac{1}{\pi} \arctan\left(\frac{x-\alpha}{\lambda}\right) + \frac{1}{2} \tag{8.64}$$

Fonction caractéristique.

$$Z(\omega) = e^{i\,\alpha\omega - |\lambda\omega|}\ . \tag{8.65}$$

Caractéristiques numériques de la loi de Cauchy.

Moments. Cette loi ne possède aucun moment et n'a donc ni moyenne ni variance.

Médiane et mode. La médiane de la loi de Cauchy $x_{0.5}$ est égale à α. C'est également la valeur de son mode.

8.3 Lois à plusieurs variables.

8.3.1 Loi multinomiale.

Un vecteur aléatoire $\boldsymbol{X} = (X_1, X_2, \ldots, X_k)$ à valeurs entières, admet la loi multinomiale de paramètres $n; p_1, p_2, \ldots, p_k$ si :

$$\boxed{\Pr\{\boldsymbol{X} = \boldsymbol{m}\} = \frac{n!}{m_1!m_2!\ldots m_k!} p_1^{m_1} p_2^{m_2} \ldots p_k^{m_k}} \tag{8.66}$$

$$\Pr\{\boldsymbol{X} = \boldsymbol{m}\} \equiv \Pr\{X_1 = m_1, X_2 = m_2, \ldots, X_k = m_k\}$$

Les paramètres p_i sont tels que : $0 \leq p_i \leq 1 \quad \sum_{i=1}^{k} p_i = 1$. L'ensemble des valeurs possibles $\boldsymbol{m} = (m_1, m_2, \ldots, m_k)$ de la variable aléatoire $\boldsymbol{X}$, est tel que : $\sum_{i=1}^{k} m_i = n$.

La loi multinomiale est une généralisation de la loi binomiale au cas où il y a plus de deux issues à une expérience aléatoire. Cette loi donne la probabilité d'obtenir m_i résultats de classe i parmi k classes, au cours de n épreuves indépendantes, lorsque la probabilité d'obtenir la classe i est p_i. Cette loi est, par exemple, utile dans l'étude des méthodes de ré-échantillonnage dites méthodes *bootstrap*.

Caractéristiques numériques de la loi multinomiale.

Moyenne. C'est un vecteur colonne de format $(k, 1)$:

$$\mathrm{E}\{\boldsymbol{X}\} = \begin{pmatrix} np_1 \\ np_2 \\ \vdots \\ np_k \end{pmatrix} \tag{8.67}$$

Matrice des variances-covariances. C'est une matrice (k, k) dont les valeurs sont données par l'expression :

$$\boldsymbol{V} = \begin{pmatrix} np_1(1-p_1) & -np_1p_2 & \cdots & -np_1p_k \\ -np_2p_1 & np_2(1-p_2) & \cdots & -np_2p_k \\ \vdots & \vdots & \ddots & \vdots \\ -np_kp_1 & -np_kp_2 & \cdots & np_k(1-p_k) \end{pmatrix} \tag{8.68}$$

Coefficients de corrélation.

$$\rho_{ij} = -\sqrt{\frac{p_i p_j}{(1-p_i)(1-p_j)}} \tag{8.69}$$

8.4 Bibliographie.

A titre d'exercices, on donne ci-dessous d'autres lois ; certaines sont extraites du chapitre 6 de l'ouvrage de Koroliouk (1983) [44]. Il existe un catalogue plus complet que l'on trouvera dans les ouvrages de Johnson *et al.* [36, 37] et [38].

8.5 Exercices et problèmes

Exercice 8.1. *Loi triangulaire de Simpson.* Soient X_1 et X_2 deux variables aléatoires indépendantes suivant la loi uniforme sur l'intervalle $[\frac{a}{2}, \frac{b}{2}]$. Montrer que la variable aléatoire $Y = X_1 + X_2$ possède une densité de probabilité $f(y)$ donnée par l'expression :

$$f(y) = \begin{cases} \dfrac{2}{b-a} - \dfrac{2}{(b-a)^2}|a+b-2y|\,, & \text{si } y \in [a,b]\,; \\ 0\,, & \text{si } y \notin [a,b]\,. \end{cases} \tag{8.70}$$

Montrer que les moments de la loi de Y sont donnés par :

$$\mathrm{E}\{Y^k\} = \frac{4}{(b-a)^2(k+1)(k+2)}\Big[a^{k+2} + b^{k+2} - 2\Big(\frac{a+b}{2}\Big)^{k+2}\Big]\,, \tag{8.71}$$

et que la variance de Y vaut :

$$\mathrm{Var}(Y) = \frac{(b-a)^2}{24}\,. \tag{8.72}$$

Exercice 8.2. *Loi bêta de type II.* Soit X une variable aléatoire suivant une loi bêta de paramètres α et β. Par définition la variable aléatoire $Y = \dfrac{X}{1-X}$ suit une loi bêta dite de type II. Démontrer que la densité de probabilité $f(y)$ de Y vaut :

$$f(y) = \begin{cases} \dfrac{1}{B(\alpha,\beta)}\,\dfrac{y^{\alpha-1}}{(1+y)^{\alpha+\beta}}\,, & \text{si } y \geq 0\,; \\ 0\,, & \text{si } y < 0\,. \end{cases} \tag{8.73}$$

Montrer que la moyenne et la variance de Y sont donnés par les expressions :

$$\mathrm{E}\,\{Y\} = \frac{\alpha}{\beta-1}, \qquad \mathrm{Var}(Y) = \frac{\alpha(\alpha+\beta-1)}{(\beta-1)^2(\beta-2)}\,\cdot \tag{8.74}$$

Montrer que si la variable aléatoire Y suit une loi bêta de type II de paramètres α, β, alors la variable aléatoire $\dfrac{\beta}{\alpha}Y$ suit une loi de Fisher de paramètres $2\alpha, 2\beta$.

Exercice 8.3. *Loi du χ.* Soit X une variable aléatoire suivant la loi du χ^2 à n degrés de liberté. Montrer que la variable aléatoire $Y = \sqrt{X}$ possède une densité de probabilité donnée par l'expression :

$$f(y) = \begin{cases} \dfrac{1}{2^{\frac{n}{2}-1}\Gamma(\frac{n}{2})}y^{n-1}\exp\{-\frac{1}{2}y^2\}\,, & \text{si } y > 0\,; \\ 0\,, & \text{si } y \leq 0\,. \end{cases} \tag{8.75}$$

La loi de Y porte le nom de loi du χ, pour $n = 2$ il s'agit de la loi de Rayleigh et pour $n = 3$ de la loi de Maxwell décrivant la vitesse des molécules d'un gaz parfait.

Montrer que les moments de la loi de Y sont égaux à :

$$\mathrm{E}\{Y^k\} = 2^{\frac{k}{2}} \frac{\Gamma(\frac{n+k}{2})}{\Gamma(\frac{n}{2})}, \tag{8.76}$$

et que la variance de Y vaut :

$$\mathrm{Var}(Y) = n - 2\left[\frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})}\right]^2. \tag{8.77}$$

Chapitre 9

Flux d'événements.

Les flux d'événements ou plus simplement, les flux, sont formés d'événements arrivant les uns après les autres, séparés par des intervalles de temps aléatoires T_i, et susceptibles d'être détectés par un dispositif quelconque. La branche de la théorie des probabilités concernant les flux est la théorie des files d'attente. Le schéma général comprend un flux d'entrée se présentant pour être servi devant n canaux. Les demandeurs servis constituent un flux de sortie et les demandeurs non-servis constituent un flux de demandes rejetées ou forment une file d'attente.

Les appels téléphoniques, l'apparition de pannes, les requêtes d'écriture ou de lecture sur un disque magnétique d'ordinateur ou encore les queues devant les guichets peuvent tous être traités suivant ce modèle. Pour le domaine qui nous concerne, les événements du flux d'entrée pourront être des photons se présentant devant un détecteur et le flux de sortie sera le signal enregistré. Certains photons peuvent ne pas être détectés si, par exemple, le détecteur est saturé.

La figure 9.1 représente schématiquement un flux d'événements. On repère les événements constituant un flux soit par les temps ξ_n auxquels ils apparaissent, soit par les intervalles de temps $T_n = \xi_n - \xi_{n-1}$ entre deux événements successifs. Tournons-nous à présent vers les caractéristiques probabilistes des flux.

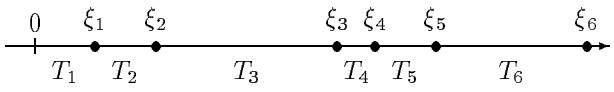


FIG. 9.1: *Représentation schématique d'un flux d'événements. Les temps d'arrivée des événements sont les variables aléatoires ξ_i et les intervalles de temps séparant les événements sont les variables aléatoires T_i. On a $T_1 = \xi_1$ et pour $i \geq 2$, $T_i = \xi_i - \xi_{i-1}$.*

9.1 Les flux simples ou de Poisson.

On s'intéresse le plus souvent à des flux possédant les propriétés suivantes.

1. Flux **stationnaire**. Un flux est stationnaire si la probabilité pour qu'un nombre quelconque d'événements apparaisse dans l'intervalle de temps $(t, t+\Delta t)$, ne dépend que de Δt et non pas du temps t. Si N désigne la variable aléatoire attachée au nombre d'événements susceptibles d'apparaitre dans le temps Δt, on aura :

$$\Pr\{N = n\} = p_n(\Delta t)\,, \tag{9.1}$$

où p_n désigne une quantité qui ne dépend du temps que par l'intermédiaire de Δt.

2. Flux **ordinaire**. Un flux est ordinaire si la probabilité pour qu'il apparaisse plus d'un événement dans le temps Δt est beaucoup plus petite que la probabilité pour qu'il apparaisse un seul événement, soit avec les notations précédentes :

$$\lim_{\Delta t \to 0} \frac{\Pr\{N > 1\}}{\Pr\{N = 1\}} = 0\ . \tag{9.2}$$

Cela implique que les événements n'arrivent pas par couples, triplets, ou suivant tout autre groupement.

3. Flux **sans post-action**. Un flux est sans post-action si les variables aléatoires T_i sont indépendantes les unes des autres et si de plus le flux est sans mémoire. Un flux est sans mémoire si la loi gouvernant les intervalles de temps entre événements obéit à la relation :

$$\forall i, \quad \Pr\{T_i > s + t | T_i > s\} = \Pr\{T_i > t\}\ . \tag{9.3}$$

Cela veut dire que, si l'événement ne s'est pas produit pendant le temps s, la probabilité pour qu'il ne se produise pas pendant un temps t supplémentaire est indépendante du temps précédent s durant lequel il ne s'est pas produit.

Un flux d'événements possédant ces trois propriétés est appelé flux simple ou flux de Poisson.

9.1.1 Loi gouvernant les intervalles de temps T_i.

Nous allons montrer que la seule loi remplissant la condition d'absence de mémoire est la loi exponentielle. La démonstration s'appuie sur celle de Rényi [62] Chapitre 3 §13.

Démonstration. Démontrons pour commencer que si la condition (9.3) est satisfaite, alors la loi suivie par T_i est exponentielle.

Soit $F(t)$ la fonction de répartition de la loi cherchée et posons $G(t) = 1 - F(t)$. On a :

$$G(t) = 1 - F(t) = 1 - \Pr\{T_i < t\} = \Pr\{T_i \geq t\}\ . \tag{9.4}$$

Notons que la variable t désigne maintenant un intervalle de temps entre deux événements. La probabilité de l'événement $\{T_i = t\}$ étant nulle on a aussi $G(t) = \Pr\{T_i > t\}$. Considérons la quantité $G(s+t) = \Pr\{T_i > s+t\}$. On a par définition de la probabilité conditionnelle :

$$\Pr\{T_i > s+t\} = \Pr\{T_i > s\} \Pr\{T_i > s+t | T_i > s\} \ . \tag{9.5}$$

Mais d'après la propriété (9.3), il vient :

$$\Pr\{T_i > s+t\} = \Pr\{T_i > s\} \Pr\{T_i > t\} \ , \tag{9.6}$$

ce qui s'écrit à l'aide des fonctions G : $G(s+t) = G(s)G(t)$. Maintenant, en choisissant $s = t$ puis $s = 2t$, on obtient $G(2t) = G^2(t)$ et $G(3t) = G^3(t)$, et par récurrence pour tout entier $n \in \mathbb{N}, G(nt) = G^n(t)$. Si l'on pose $s = nt$ on obtient $G(s)^{1/n} = G(s/n)$ où s et n sont quelconques. Choisissons $s = mt$, il vient $G(\frac{m}{n}t) = G(t)^{m/n}$. En posant $t = 1$ et $m/n = r \in \mathbb{Q}^+$ dans l'équation précédente, on obtient pour tout rationnel positif r : $G(r) = G(1)^r$. La quantité $G(1)$ désigne la probabilité pour que l'on n'observe pas d'événement dans l'unité de temps. Comme $G(1) \leq 1$, on peut poser $G(1) = e^{-\lambda}$ et l'on a alors pour tout t rationnel positif $G(t) = e^{-\lambda t}$.

La fonction G étant monotone, l'encadrement des réels par les rationnels sur l'axe des t se retrouve bijectivement pour $G(t)$ et l'on peut poser par continuité pour $t \in \mathbb{R}$:

$$F(t) = 1 - G(t) = 1 - e^{-\lambda t} \ . \tag{9.7}$$

Remarquons que seule l'hypothèse d'absence de mémoire a été utilisée pour déduire ce résultat. Afin de trouver le sens qu'il convient de donner à la valeur λ, nous allons maintenant faire appel aux hypothèses 1 et 2. Calculons la probabilité pour qu'il y ait au moins un événement dans l'intervalle de temps Δt :

$$\Pr\{T_i \leq \Delta t\} = F(\Delta t)$$

$$F(\Delta t) = 1 - e^{-\lambda \Delta t} = \lambda \Delta t + o((\Delta t)^2) \ . \tag{9.8}$$

D'après l'hypothèse de stationnarité, cette quantité ne dépend que de Δt. Cela montre que la quantité λ est une constante, indépendante du temps. Par ailleurs quand $\Delta t \to 0$, d'après l'hypothèse numéro 2 il n'y a que deux possiblités : soit il n'y a pas d'événement dans Δt, soit on n'en observe qu'un. On observe donc en moyenne dans Δt :

$$\mathrm{E}\{N(\Delta t)\} = 0 \times (1 - \lambda \Delta t) + 1 \times \lambda \Delta t = \lambda \Delta t \ ,$$

et dans cette expression $N(\Delta t)$ est une variable aléatoire égale au nombre d'événements dans l'intervalle de temps Δt. Il vient alors :

$$\lambda = \lim_{\Delta t \to \infty} \frac{\mathrm{E}\{N(\Delta t)\}}{\Delta t} \tag{9.9}$$

La constante λ est donc le nombre moyen d'événements par unité de temps. Nous avons ainsi démontré que les intervalles de temps séparant les événements constituant un flux de Poisson suivaient une loi exponentielle, de fonction de répartition et de densité de probabilité données par :

$$F(t) = (1 - e^{-\lambda t})H(t), \quad f(t) = (\lambda e^{-\lambda t})H(t), \tag{9.10}$$

où $H(t)$ est la distribution de Heaviside.

Il reste à démontrer la propriété réciproque, c'est-à-dire que la loi exponentielle répond à la condition (9.3). En effet si la loi suivie par T_i est exponentielle on a d'après la définition des probabilités conditionnelles :

$$\Pr\{T_i > s+t | T_i > s\} = \frac{\Pr\{T_i > s+t, T_i > s\}}{\Pr\{T_i > s\}}$$

Le numérateur de la fraction du second membre se simplifie, car si la condition $T_i > s+t$ est satisfaite, alors la condition $T_i > s$ est nécessairement satisfaite. Il vient donc :

$$\begin{aligned}\Pr\{T_i > s+t | T_i > s\} &= \frac{\Pr\{T_i > s+t\}}{\Pr\{T_i > s\}} \\ &= \frac{1-(1-e^{-\lambda(t+s)})}{1-(1-e^{-\lambda s})} \\ &= e^{-\lambda t} = \Pr\{T_i > t\} \ ,\end{aligned}$$

ce qui correspond bien à la définition de la propriété d'absence de mémoire. □

9.1.2 Lois gouvernant les temps d'arrivée ξ_i des événements.

Il y a identité entre ξ_1 et T_1. La variable aléatoire ξ_1 suit donc elle aussi une loi exponentielle de paramètre λ. Pour calculer la loi suivie par ξ_2 remarquons que c'est la somme de deux variables aléatoires par hypothèse indépendantes :

$$\xi_2 = T_2 + T_1 \ .$$

La densité de probabilité f_2 de ξ_2 est alors la convolution de deux densités exponentielles. Il vient :

$$\begin{aligned}f_2(t) &= \int_{-\infty}^{\infty} \lambda e^{-\lambda u} H(u) \lambda e^{-\lambda(t-u)} H(t-u)\, du \\ &= \lambda^2 e^{-\lambda t} \int_0^t du \\ f_2(t) &= \lambda^2 t e^{-\lambda t} \ .\end{aligned}$$

Le temps d'arrivée du deuxième événement suit donc la loi d'Erlang d'ordre 2. On démontrerait facilement par récurrence que le temps d'arrivée ξ_n du n^{e} événement suit une loi d'Erlang d'ordre n :

$$f_n(t) = \frac{\lambda^n}{(n-1)!} t^{n-1} e^{-\lambda t} \ . \tag{9.11}$$

9.1.3 Loi gouvernant le nombre d'événements observés dans un intervalle de temps donné T.

Le nombre d'événements observés dans l'intervalle de temps T est une variable aléatoire que nous noterons N. C'est a priori une fonction de T, $N = N(T)$. Calculons la probabilité pour obtenir exactement n événements dans T. Pour cela il faut que $\xi_n \leq T$ ce qui assure que l'on observe au moins n événements dans T, mais il faut aussi que $\xi_{n+1} > T$ de façon à en observer exactement n (voir figure 9.2). Par définition $\xi_{n+1} = \xi_n + T_{n+1}$, les variables aléatoires ξ_n et T_{n+1} sont indépendantes, ce que l'on montre facilement par changement de variable sur l'ensemble des $n+1$ variables indépendantes T_i. La densité de probabilité du couple (T_{n+1}, ξ_n) vaut donc :

$$f_{T_{n+1},\xi_n}(t_1, t_2) = \lambda e^{-\lambda t_1} \frac{\lambda^n}{(n-1)!} t_2^{n-1} e^{-\lambda t_2} \ . \tag{9.12}$$

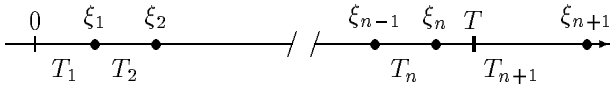


FIG. 9.2: *Flux d'événements correspondant au cas où l'on a exactement n événements dans l'intervalle de temps T.*

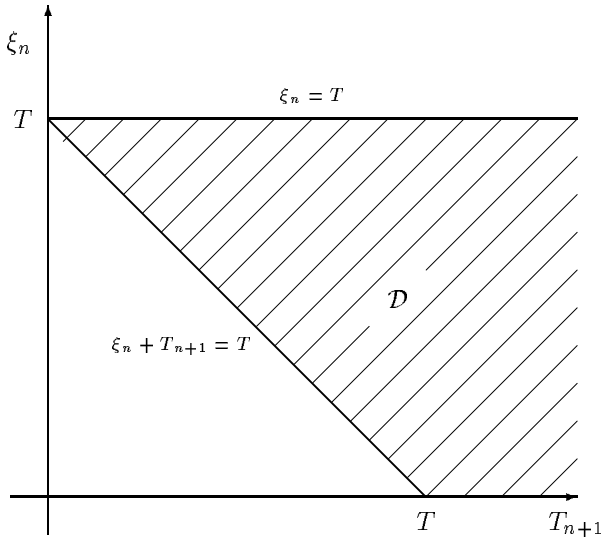


FIG. 9.3: *Domaine d'intégration $\mathcal{D}$, satisfaisant les conditions $\xi_n \leq T$ et $\xi_n + T_n > T$. Ces conditions correspondent aux cas où l'on observe exactement n événements dans le temps T.*

Le domaine $\mathcal{D}$ du plan (T_{n+1}, ξ_n) satisfaisant aux conditions $\xi_n \leq T$ et $\xi_{n+1} = \xi_n + T_{n+1} > T$ est illustré par la figure 9.3. Il est maintenant facile de calculer la probabilité p_n associée à l'événement $\{N = n\}$. En effet :

$$p_n = \Pr\{N = n\} = \iint_{\mathcal{D}} f_{T_{n+1}\xi_n}(t_1, t_2)\, dt_1 dt_2$$

$$\begin{aligned}
p_n &= \frac{\lambda^{n+1}}{(n-1)!} \iint_{\mathcal{D}} e^{-\lambda t_1} t_2^{n-1} e^{-\lambda t_2}\, dt_1 dt_2 \\
&= \frac{\lambda^{n+1}}{(n-1)!} \int_0^T t_2^{n-1} e^{-\lambda t_2}\, dt_2 \int_{T-t_2}^{\infty} e^{-\lambda t_1}\, dt_1 \\
&= \frac{\lambda^{n+1}}{(n-1)!} \int_0^T t_2^{n-1} e^{-\lambda t_2} \frac{1}{\lambda} e^{-\lambda(T-t_2)}\, dt_2 \\
&= \frac{\lambda^n}{(n-1)!} e^{-\lambda T} \int_0^T t_2^{n-1}\, dt_2 \\
&= \frac{\lambda^n}{(n-1)!} e^{-\lambda T} \frac{T^n}{n},
\end{aligned}$$

d'où la solution cherchée :

$$p_n = \Pr\{N = n\} = \frac{(\lambda T)^n}{n!} e^{-\lambda T} \ . \tag{9.13}$$

Le nombre d'événements susceptibles d'apparaître dans le temps T suit une loi de Poisson de paramètre $\mu = \lambda T$, et par conséquent de moyenne $\mathrm{E}\{N\} = \lambda T$ et de variance $\mathrm{Var}(N) = \lambda T$.

▶**Exemple 9.1.** *Bruit de photons.* Si le modèle corpusculaire de la lumière peut s'appliquer, on peut raisonnablement identifier à un flux de Poisson, le flux de photons enregistré par un détecteur parfait quelconque. Supposons que la source lumineuse émette des photons à un taux de λ photons par unité de temps. Supposons de plus que le détecteur soit effectivement parfait, c'est-à-dire qu'il enregistre les impacts des photons sans pertes ni délais aléatoires et qu'il n'introduit pas d'autres sources de bruit. Les calculs précédents montrent que les temps T séparant les impacts des photons doivent suivre la loi exponentielle de densité : $f(t) = \lambda e^{-\lambda t}$, de moyenne $\mathrm{E}\{T\} = 1/\lambda$ et de variance $\mathrm{Var}(T) = 1/\lambda^2$. Le nombre N d'impacts susceptibles d'être enregistrés dans un intervalle de temps Δt suit la loi de Poisson de moyenne $\mathrm{E}\{N\} = \lambda \Delta t$ et de variance $\mathrm{Var}(N) = \lambda \Delta t$.

Si le détecteur compte les photons par tranche de temps Δt, le signal enregistré pourra ressembler à celui de la figure 9.4.

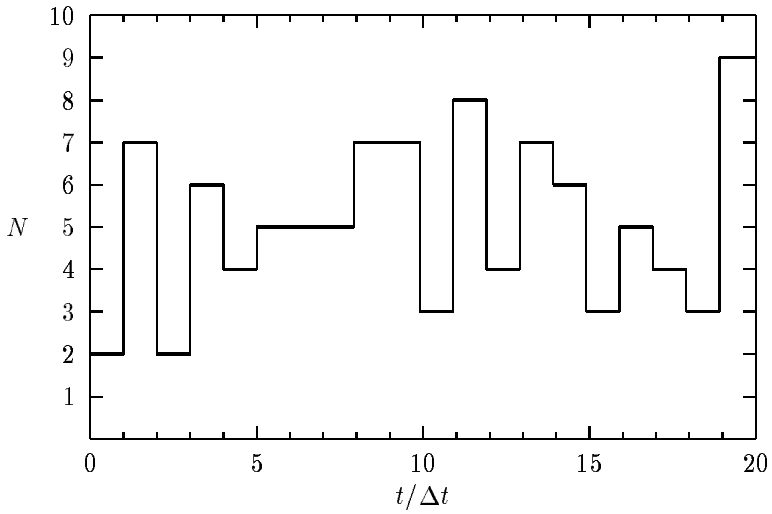


FIG. 9.4: *Exemple du nombre de photons enregistrés par intervalle de temps Δt, quand on reçoit en moyenne 5 photons dans cet intervalle Δt.*

La moyenne $\lambda \Delta t$ est en fait le signal que l'on cherche à détecter et le bruit parasite peut être quantifié par l'écart type $\sqrt{\lambda \Delta t}$, de sorte que le rapport signal sur bruit SN, qui est un critère de qualité de l'observation, est donné par l'expression :

$$SN = \frac{\lambda \Delta t}{\sqrt{\lambda \Delta t}} = \sqrt{\lambda \Delta t} \ . \tag{9.14}$$

Le rapport signal sur bruit augmente donc comme $\sqrt{\Delta t}$ où Δt est le temps d'observation du signal.

9.1.4 Propriété réciproque.

Nous allons maintenant établir la propriété réciproque, c'est-à-dire que si un flux est tel que le nombre $N(T)$ d'événements observé dans un intervalle de temps T quelconque suit la loi de Poisson, et si les variables aléatoires $N(T_i)$ sont indépendantes pour tous les intervalles T_i disjoints, alors le flux est un flux de Poisson.

Par hypothèse, la variable aléatoire $N(T)$ suit la loi de Poisson :

$$\Pr\{N(T) = n\} = \frac{(\lambda T)^n}{n!} e^{-\lambda T} , \tag{9.15}$$

où λ est une constante dont la signification est le nombre moyen d'événements par unité de temps. Montrons que ces hypothèses remplissent les conditions requises par le flux de Poisson.

1. Stationnarité. Par définition la quantité $\Pr\{N(\Delta T) = n\}$ ne dépend pas du temps si λ ne dépend pas lui-même du temps.

2. Flux ordinaire. On a les limites suivantes :

$$\begin{aligned}
\lim_{\Delta t \to 0} \Pr\{N(\Delta t) = 0\} &= 1 - \lambda \Delta t \\
\lim_{\Delta t \to 0} \Pr\{N(\Delta t) = 1\} &= \lambda \Delta t - \lambda^2 (\Delta t)^2 \\
\lim_{\Delta t \to 0} \Pr\{N(\Delta t) > 1\} &= \lambda^2 (\Delta t)^2 ,
\end{aligned}$$

   ce qui montre bien que $\lim_{\Delta t \to 0} \Pr\{N(\Delta t) > 1\} / \Pr\{N(\Delta t) = 1\} = 0$, et que le flux est ordinaire.

3. Flux sans post-action. Montrons tout d'abord que les intervalles de temps T_i entre événements suivent la loi exponentielle. En effet il vient :

$$\Pr\{T_i \geq t\} = \Pr\{N(t) = 0\} = \frac{(\lambda t)^0}{0!} e^{-\lambda t} = e^{-\lambda t} ,$$

   d'où la fonction de répartition de T_i :

$$F(T_i) = 1 - \Pr\{T_i \geq t\} = 1 - e^{-\lambda t} , \tag{9.16}$$

   ce qui montre que T_i suit la loi exponentielle et cette loi est bien sans mémoire d'après la démonstration du chapitre 9.1.1. La propriété d'indépendance des T_i découle de celle de l'indépendance des $N(T_i)$; montrons-le rapidement pour un couple disjoint T_i, T_j. On a :

$$\begin{aligned}
\Pr\{T_i < s, T_j < t\} &= \Pr\{N(s) > 0, N(t) > 0\} \\
&= \Pr\{N(s) > 0\} \Pr\{N(t) > 0\} \\
&= \Pr\{T_i < s\} \Pr\{T_j < t\} .
\end{aligned}$$

Ce dernier point achève la démonstration, et le flux est donc bien un flux de Poisson.

9.2 Flux de Poisson non-stationnaire.

Nous allons abandonner, dans cette partie, la propriété d'invariance du flux par translation de l'origine des temps. Il est donc maintenant nécessaire de s'intéresser au nombre d'événements susceptibles de survenir dans l'intervalle de temps t_1, t_2. Soit $N(t_1, t_2)$ la variable aléatoire associée à ce nombre. Nous dirons qu'un flux ordinaire est un flux de Poisson « *non-stationnaire* » s'il possède d'abord les propriétés suivantes.

1. Indépendance. Si $t_1 < t_2 \leq t_3 < t_4$, alors les variables aléatoires $N(t_1, t_2)$ et $N(t_3, t_4)$ sont indépendantes. C'est la propriété d'indépendance sur des intervalles de temps disjoints que nous avons déjà vue dans le cas stationnaire.

2. Existence d'une densité instantanée. Quel que soit t, le nombre moyen d'événements dans l'intervalle $t, t + \Delta t$ divisé par Δt, tend vers une limite quand Δt tend vers zéro. Plus précisément :

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{\mathrm{E}\left\{N(t, t + \Delta t)\right\}}{\Delta t} \quad . \tag{9.17}$$

   La quantité $\lambda(t)$ est appelé la « *densité instantanée* » du flux d'événements. C'est le taux auquel arrivent les événements constituant le flux. Un flux de Poisson non-stationnaire est donc un flux à taux variable.

9.2.1 L'horloge stroboscopique.

Il existe un moyen simple de ramener un flux à taux variable à un flux à taux fixe. Il suffit pour cela de changer l'horloge qui enregistre les dates d'arrivées des événements et de considérer une nouvelle horloge qui accélère lorsque le taux $\lambda(t)$ augmente et qui ralentit lorsqu'il diminue. La nouvelle horloge doit être réglée de façon telle que le taux d'événements, rapporté à ce nouveau temps, apparaisse constant. Plus précisément, si on considère le temps Θ définit ainsi :

$$\Theta = \int_0^t \lambda(u)\, du \,, \tag{9.18}$$

alors les événements arrivent à un taux constant égal à l'unité.

Démonstration. Si $\lambda'(\Theta)$ désigne le nouveau taux d'arrivée des événements par rapport à Θ, on a par définition :

$$\lambda'(\Theta) = \lim_{\Delta\Theta \to 0} \frac{\mathrm{E}\left\{N(\Theta, \Theta + \Delta\Theta)\right\}}{\Delta\Theta} \,.$$

Si $\lambda(t)$ n'est pas nul (ni infini) la fonction (9.18) qui fait passer de t à Θ est inversible et continue pour tout t. D'après le théorème de la fonction inverse, le nombre d'événements dans l'intervalle $(\Theta, \Theta + \Delta\Theta)$ est égal à celui trouvé dans l'intervalle $(t, t + \Delta t)$, soit : $N(\Theta, \Theta + \Delta\Theta) = N(t, t + \Delta t)$. Par ailleurs on a $\Delta\Theta = \lambda(t)\Delta t + O(\Delta t^2)$, il vient alors :

$$\begin{aligned} \lambda'(\Theta) &= \lim_{\Delta t \to 0} \frac{\mathrm{E}\{N(t, t + \Delta t)\}}{\Delta t} \frac{\Delta t}{\Delta\Theta} \,, \\ &= \lambda(t)\Big(\frac{d\Theta}{dt}\Big)^{-1} = \lambda(t)(\lambda(t))^{-1} = 1 \,. \end{aligned}$$

Le flux en Θ est donc constant égal à un événement par unité de Θ. □

3. Pour que le flux en Θ soit de Poisson il faut, de plus, postuler l'absence de mémoire, c'est-à-dire :

$$\Pr\{\Theta_i > s + t | \Theta_i > s\} = \Pr\{\Theta_i > t\} , \tag{9.19}$$

où les Θ_i désignent les temps d'arrivées des événements suivant l'horloge Θ.

Il faut donc rajouter cette condition 3 aux conditions 1 et 2 pour qu'un flux à λ variable soit un flux de Poisson non-stationnaire.

9.2.2 Loi du nombre d'événements dans un intervalle t_1, t_2

Dans les conditions précédentes, on trouve facilement que le nombre d'événements contenus dans l'intervalle t_1, t_2 suit une loi de Poisson de paramètre $\mu = \int_{t_1}^{t_2} \lambda(u)\, du$. En effet, on a :

$$\begin{aligned} \Pr\{N(t_1,t_2) = n\} = \Pr\{N(\Theta_1,\Theta_2) = n\} &= \frac{(\Theta_2 - \Theta_1)^n}{n!} \exp\{-(\Theta_2 - \Theta_1)\} , \\ &= \frac{\mu^n}{n!} e^{-\mu}, \quad \text{avec} \quad \mu = \int_{t_1}^{t_2} \lambda(t)\, dt . \end{aligned} \tag{9.20}$$

9.2.3 Loi suivie par l'intervalle de temps séparant deux événements.

Le temps t_0 étant donné, la loi suivie par le temps t écoulé avant l'apparition d'un événement est donnée par $1 - e^{-(\Theta - \Theta_0)}$, d'où on tire immédiatement la fonction de répartition :

$$F_{t_0}(t) = 1 - e^{-\int_{t_0}^{t_0+t} \lambda(u)\, du} , \tag{9.21}$$

et la densité de probabilité :

$$f_{t_0}(t) = \lambda(t_0 + t) e^{-\int_{t_0}^{t_0+t} \lambda(u)\, du} . \tag{9.22}$$

Comme il se doit, on retrouve bien la loi exponentielle dans le cas où $\lambda(t) = \lambda = Cste$.

▶**Exemple 9.2.** *Source modulée sinusoïdalement.* Supposons que l'on observe une source de lumière dont la densité instantanée de photons est :

$$\lambda(t) = \lambda_0 + \lambda_1 \sin \omega t, \quad \lambda_1 \leq \lambda_0 .$$

En appliquant la formule (9.22) on trouve :

$$f_{t_0}(t) = [\lambda_0 + \lambda_1 \sin \omega(t_0 + t)]\, e^{-\lambda_0 t - \frac{2\lambda_1}{\omega} \sin \omega(t_0 + \frac{t}{2}) \sin \omega \frac{t}{2}} . \tag{9.23}$$

La figure 9.5 représente cette fonction pour $t_0 = 0$, $\lambda_0 = \lambda_1 = 1$ et pour $\omega = 2\pi$.

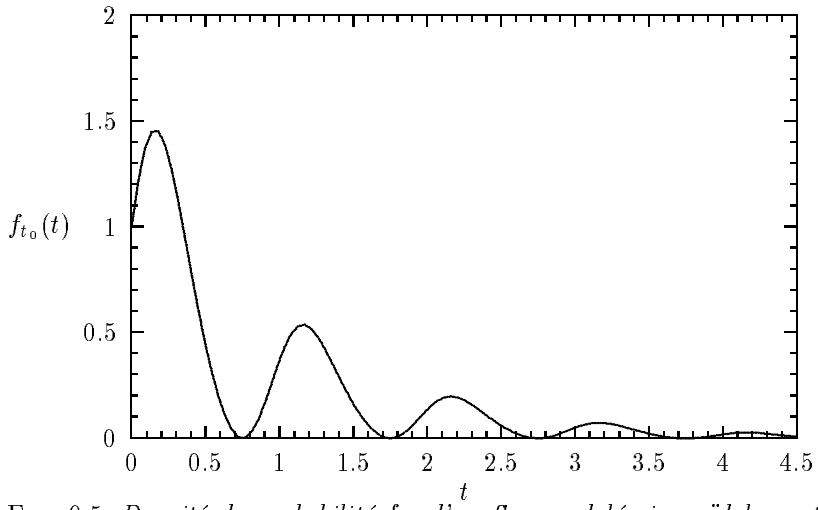


FIG. 9.5: *Densité de probabilité f_{t_0} d'un flux modulé sinusoïdalement, suivant la densité d'événements $\lambda = 1 + \sin 2\pi t$. On a tracé la densité de probabilité f_{t_0} pour les temps t comptés à partir de l'origine $t_0 = 0$.*

9.3 Superposition de flux.

9.3.1 Définition.

Considérons une série d'événements issus d'un flux F_1 apparus aux temps $(x_1^1, x_2^1, \ldots)$ et une autre série d'événements issus d'un flux F_2, apparus aux temps $(x_1^2, x_2^2, \ldots)$, les flux F_1 et F_2 ayant la même origine des temps. Reportons ces temps sur le même axe et considérons-les comme les temps d'apparition d'une nouvelle série d'événements. Cette procédure est illustrée par la figure 9.6. Supposons que l'on fasse cette opération pour toutes les apparitions

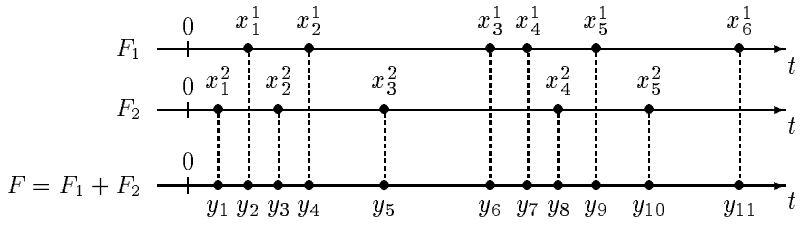


FIG. 9.6: *Représentation schématique de la superposition, ou somme de deux flux.*

possibles des événements constituant les flux F_1 et F_2 ; on obtiendra ainsi toutes les apparitions possibles des événements d'un nouveau flux F que l'on définira comme étant la somme des flux F_1 et F_2. Cela étant défini, le problème à résoudre maintenant est de connaître les caractéristiques probabilistes de F, connaissant celles de F_1 et F_2.

9.3.2 Flux indépendants.

Nous dirons que deux flux F_1 et F_2 sont indépendants, si pour $t_{11} < t_{21}$ et $t_{12} < t_{22}$, les variables aléatoires $N_1(t_{11}, t_{21})$ et $N_2(t_{12}, t_{22})$ associées respectivement aux flux F_1 et F_2, sont indépendantes. On généralise sans peine cette notion à plusieurs flux, et la condition nécessaire et suffisante pour que n flux soient indépendants est :

$$\Pr\{N_1(t_{11}, t_{21}), N_2(t_{12}, t_{22}), \ldots, N_n(t_{1n}, t_{2n})\} = \prod_{i=1}^{n} \Pr\{N_i(t_{1i}, t_{2i})\}, \quad (9.24)$$

pour tous les $t_{1i} < t_{2i}$. Rappelons de nouveau que si des flux sont deux à deux indépendants, ils ne sont pas nécessairement indépendants.

9.3.3 Superposition de flux de Poisson.

La superposition de deux flux de Poisson indépendants, respectivement de paramètres λ_1 et λ_2, est un flux de Poisson de paramètre $\lambda = \lambda_1 + \lambda_2$.

Montrons que le flux $F = F_1 + F_2$ est bien de Poisson, c'est-à-dire qu'il est ordinaire, que le nombre d'événements N dans l'intervalle de temps $t_1 < t_2$ suit la loi de Poisson et que pour deux intervalles de temps disjoints $t_1 < t_2$ et $t_3 < t_4$, les variables aléatoires associées N_1 et N_2 correspondant aux nombres d'événements susceptibles d'y être observées sont des variables aléatoires indépendantes.

1. Flux ordinaire. La superposition de deux flux ordinaires est ordinaire.

2. Loi de Poisson. La somme de deux variables aléatoires indépendantes suivant la loi de Poisson suit également une loi de Poisson, le paramètre de la somme étant égal à la somme des paramètres.

3. Indépendance. L'indépendance des variables aléatoires « comptant » le nombre d'événements sur des intervalles de temps disjoints, résulte directement de cette propriété dont jouissent les flux de Poisson et de l'indépendance des flux entre eux.

Notons que nous n'avons pas supposé que les flux étaient stationnaires, et donc la somme de deux flux de Poisson non-stationnaires est un flux de Poisson a priori non-stationnaire de paramètre $\lambda(t) = \lambda_1(t) + \lambda_2(t)$.

9.3.4 Tendance vers le flux de Poisson.

On montre, sous des hypothèses assez générales, que la somme de n flux indépendants ordinaires et stationnaires, tend vers un flux de Poisson quand $n \to \infty$. Le flux de Poisson joue vis-à-vis de la somme de flux indépendants, le même rôle que la loi normale joue vis-à-vis de la somme de variables aléatoires indépendantes et de variances finies. Remarquons que les flux dont on fait la somme peuvent être à post-action quelconque, et donc que la somme de flux indépendants tend à diluer l'effet des post-actions individuelles.

▶**Exemple 9.3.** *Source radioactive.* Une source radioactive est formée de N noyaux atomiques susceptibles d'émettre par exemple des particules α. Un noyau individuel dans un état excité ne peut émettre qu'une seule particule α. Le flux associé à ce seul noyau est nécessairement ordinaire, il est sans mémoire car la loi qui préside à la

désintégration nucléaire est la loi exponentielle ; en revanche, il n'est pas stationnaire car le nombre moyen de désintégration(s) par seconde diminue avec le temps comme la loi exponentielle. La propriété de post-action n'a pas de sens dans le cas présent.

On montre cependant que l'absence de mémoire compensant la non-stationnarité, la superposition d'un nombre $N \to \infty$ de flux associés aux noyaux, tend rapidement vers un flux de Poisson. Cet exemple ne répond pas précisément aux conditions énoncées ci-dessus, mais illustre l'effet attractif de la loi de Poisson.

9.4 Flux tamisés.

On obtient un flux « *tamisé* » en supprimant certains événements constituant le flux. Ce « *tamisage* » peut être déterministe si par exemple on supprime un point sur deux, ou aléatoire si un événement est supprimé ou non suivant l'issue d'une certaine variable aléatoire. Le nouveau flux ainsi construit est appelé « *flux tamisé* ».

9.4.1 Flux d'Erlang.

Pour obtenir un flux d'Erlang d'ordre k, on supprime k événements successifs d'un flux de Poisson et l'on conserve le $k+1$-ième ; on ne conserve donc que les événements dont l'indice est divisible par $k+1$, voir figure 9.7.

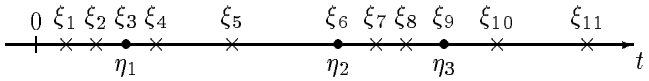


FIG. 9.7: *Tamisage déterministe d'un flux de Poisson. On ne conserve qu'un point sur 3, le flux résultant est un flux d'Erlang d'ordre 2.*

D'après ce que nous avons vu plus haut, la loi qui décrit les intervalles de temps T d'un flux d'Erlang d'ordre k est la loi d'Erlang (ou loi gamma) dont la densité de probabilité f_k est donnée par :

$$f_k(t) = \frac{\lambda^{k+1} t^k}{k!} e^{-\lambda t} \ . \tag{9.25}$$

On peut retrouver très facilement cette loi en remarquant que pour que la variable aléatoire T se trouve dans l'intervalle $t, t+dt$ avec la probabilité $f_k(t)dt$, il faut avoir k événements dans le temps t (ceux qui on été supprimés), ce qui arrive avec la probabilité $(\lambda t)^k e^{-\lambda t}/k!$, et qu'il faut avoir un événement dans le temps dt suivant, ce qui arrive avec la probabilité λdt. Compte tenu des propriétés d'indépendance du flux de Poisson :

$$f_k(t)dt = \frac{(\lambda t)^k}{k!} e^{-\lambda t} \lambda dt, \tag{9.26}$$

d'où la densité cherchée. Cette loi possède une moyenne et une variance données par les expressions :

$$\mathrm{E}\{T\} = \frac{k+1}{\lambda}, \quad \mathrm{Var}(T) = \frac{k+1}{\lambda^2} \ . \tag{9.27}$$

Remarquons, pour finir, qu'un flux d'Erlang possède une post-action car la loi suivie par les intervalles de temps entre événements n'est pas exponentielle. Cependant, deux intervalles de temps successifs étant indépendants, le flux d'Erlang possède ce que l'on appelle une post-action « limitée ».

▶**Exemple 9.4.** *Photométrie du 32-ième photon.* On observe une source en comptant les photons qu'elle émet. Sur une bande magnétique, on enregistre les temps d'arrivée de ces photons. Afin de réduire la masse des données à enregistrer, on décide de n'enregistrer qu'un point sur 32. Si la densité de photons λ peut être considérée comme constante entre deux enregistrements, la loi décrivant le temps séparant deux enregistrements est la loi d'Erlang d'ordre $k = 31$. Si l'on veut estimer λ à partir de deux enregistrements consécutifs séparés par le temps t, on prendra $\hat{\lambda} = (k+1)/t$. Il est facile de voir d'après (9.27) que $\mathrm{E}\{\hat{\lambda}\} = \lambda$ et que l'écart type de cette estimation est $\sigma = (k+1)^{\frac{1}{2}}/\lambda$. Si l'on avait conservé tous les k photons séparés par des temps t_i, on aurait probablement estimé λ soit comme la moyenne arithmétique des inverses des t_i, soit comme l'inverse de la moyenne arithmétique des t_i. Nous verrons plus loin (section 15.1) que la première estimation est moins bonne que la seconde et que cette dernière est en fait la meilleure possible telle que $\mathrm{E}\{\hat{\lambda}\} = \lambda$. Mais alors cette deuxième estimation est identique à celle trouvée en ne conservant que le $k+1$-ième photon et ainsi on ne perd rien vis-à-vis de l'estimation de λ en comprimant l'information de la façon qui vient d'être décrite. Le point essentiel est que λ doit pouvoir être considéré comme constant entre deux enregistrements consécutifs.

9.4.2 Tamisage aléatoire d'un flux de Poisson.

Dans la pratique, il arrive très souvent qu'un flux de Poisson soit tamisé de la façon suivante : on garde un événement avec une probabilité p et on le rejette avec une probabilité $(1-p)$. On procède ainsi pour chaque événement constituant le flux et cela de façon indépendante des autres événements. Le tamisage du flux est donc subordonné à une variable aléatoire de Bernoulli.

Si le flux de Poisson était au départ de paramètre λ, le flux tamisé sera également un flux de Poisson mais de paramètre $p\lambda$. C'est en effet un flux ordinaire, l'indépendance sur des intervalles de temps disjoints n'est pas modifiée par le tamisage qui est lui même indépendant, et la probabilité p_n d'observer n événements dans l'intervalle de temps t est donnée par :

$$\begin{aligned} p_n &= \sum_{k=n}^{\infty} \frac{(\lambda t)^k}{k!} e^{-\lambda t} C_k^n p^n (1-p)^{k-n} \\ &= (p\lambda t)^n e^{-\lambda t} \sum_{k=n}^{\infty} (1-p)^{k-n} C_k^n \frac{(\lambda t)^{k-n}}{k!} \\ &= \frac{(p\lambda t)^n}{n!} e^{-\lambda t} \sum_{k=n}^{\infty} (1-p)^{k-n} \frac{(\lambda t)^{k-n}}{(k-n)!} \\ &= \frac{(p\lambda t)^n}{n!} e^{-\lambda t} \sum_{k=0}^{\infty} (1-p)^{k} \frac{(\lambda t)^{k}}{k!} \\ &= \frac{(p\lambda t)^n}{n!} e^{-\lambda t} e^{\lambda t(1-p)}, \end{aligned}$$

soit finalement :

$$p_n = \frac{(p\lambda t)^n}{n!} e^{-p\lambda t}, \tag{9.28}$$

ce qui montre bien que le flux ainsi tamisé est un flux de Poisson de paramètre $p\lambda$.

▶**Exemple 9.5.** *Tamisage d'un flux de photons par une photocathode.* Un flux de photons se présente comme un flux de Poisson de paramètre λ. Les photons sont transformés en photo-électrons par une photocathode dont le rendement est de 20%. Le « choix » des photons pouvant être considéré comme aléatoire, le flux de photo-électrons sera un flux de Poisson de paramètre 0.2λ.

9.5 Flux 2D.

Un flux 2D (à deux dimensions) est formé d'événements répartis au hasard sur le plan $\mathbb{R}^2$. On représente souvent un flux 2D comme un ensemble de points dont les coordonnées cartésiennes sont des variables aléatoires, comme illustré par la figure 9.8. On peut également repérer le point représentatif d'un événe-

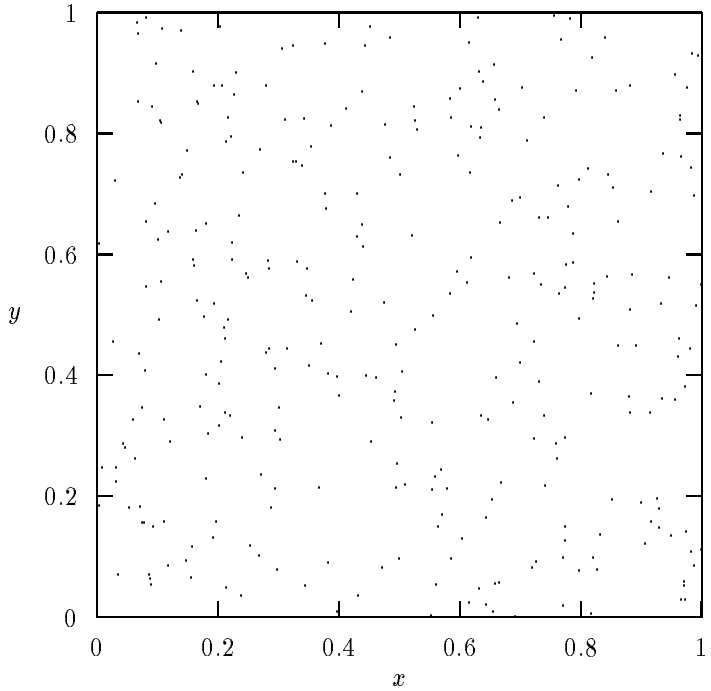


FIG. 9.8: *Portion de flux 2D vu à travers la fenêtre* $[0,1]\times[0,1]$ *de* $\mathbb{R}^2$. *Le flux représenté ici est un flux de densité uniforme de* $\lambda = 300$ *points par unité de surface.*

ment par son rayon vecteur $\boldsymbol{r}$.

9.5.1 Caractéristiques locales d'un flux 2D.

Densité en un point. On caractérise un flux par la densité d'événements en un point $\boldsymbol{r}$. Le nombre d'événements $N(\Delta S)$ contenu dans une surface ΔS est une variable aléatoire discrète qui possède une moyenne. A l'aide de cette

moyenne on introduit la densité du flux en 1 point :

$$n_1(\boldsymbol{r}) = \lim_{\Delta S \to 0} \frac{\mathrm{E}\{N|\Delta S\}}{\Delta S} . \tag{9.29}$$

Densité et fonction de corrélation en deux points. La densité en 2 points $n_2(\boldsymbol{r}_1, \boldsymbol{r}_2)$ est, de façon analogue, égale à la moyenne du nombre de points dans deux surfaces disjointes ΔS_1 et ΔS_2. Plus précisément :

$$n_2(\boldsymbol{r}_1, \boldsymbol{r}_2) = \lim_{\Delta S_1, \Delta S_2 \to 0} \frac{\mathrm{E}\{N|\Delta S_1, \Delta S_2\}}{\Delta S_1 \Delta S_2} . \tag{9.30}$$

La « *fonction de corrélation en deux points* » $\xi(\boldsymbol{r}_1, \boldsymbol{r}_2)$ est définie à l'aide de la densité en 2 points comme suit :

$$n_2(\boldsymbol{r}_1, \boldsymbol{r}_2) = n_1(\boldsymbol{r}_1) n_1(\boldsymbol{r}_2)(1 + \xi(\boldsymbol{r}_1, \boldsymbol{r}_2)) . \tag{9.31}$$

Distance aux plus proches voisins. La distance d'un point quelconque du plan $\mathbb{R}^2$ au n-ième plus proche voisin est une variable aléatoire que nous noterons D_n. Elle possède une densité de probabilité que nous noterons f_{D_n}.

9.5.2 Propriétés globales d'un flux 2D.

Comme pour les flux 1D (sur $\mathbb{R}$) on s'intéresse plus particulièrement à des flux 2D possédant certaines propriétés globales dont voici les plus usuelles.

1. Flux homogène. Un flux 2D est homogène si les événements qui le constituent sont de même nature, ou, si l'on préfère, s'ils sont indiscernables, sinon sous tous leurs aspects, du moins au regard des propriétés qui nous intéressent.

2. Flux ordinaire. Un flux 2D est ordinaire si la probabilité d'observer un événement dans une petite surface ΔS est beaucoup plus grande que celle d'observer plus d'un événement dans cette même surface. Plus précisément un flux est ordinaire si :

$$\lim_{\Delta S \to 0} \frac{\Pr\{N(\boldsymbol{r}|\Delta S) > 1\}}{\Pr\{N(\boldsymbol{r}|\Delta S) = 1\}} = 0 . \tag{9.32}$$

3. Flux uniforme. Un flux 2D est uniforme si la densité en un point $n_1(\boldsymbol{r})$ ne dépend pas de $\boldsymbol{r}$. Cette propriété est l'analogue 2D de la propriété de stationnarité d'un flux 1D.

4. Flux isotrope. Un flux 2D est isotrope si la fonction de corrélation en 2 points ne dépend que du module de la distance entre les 2 points.

5. Flux sans action à distance. Un flux 2D possède cette propriété si étant donné deux surfaces disjointes quelconques ΔS_1 et ΔS_2, le nombre d'événements N qu'elles sont susceptibles de contenir sont des variables aléatoires indépendantes. Le flux 2D sera donc sans action à distance si :

$$\begin{aligned} \Pr\{N(\boldsymbol{r}_1|\Delta S_1) = m_1, N(\boldsymbol{r}_2|\Delta S_2) = m_2\} = \\ = \Pr\{N(\boldsymbol{r}_1|\Delta S_1) = m_1)\} \Pr\{N(\boldsymbol{r}_2|\Delta S_2) = m_2\} . \end{aligned} \tag{9.33}$$

Cela implique la relation suivante entre les densités à 1 point et à 2 points :

$$n_2(\boldsymbol{r}_1, \boldsymbol{r}_2) = n_1(\boldsymbol{r}_1) n_1(\boldsymbol{r}_2), \tag{9.34}$$

et sur la fonction de corrélation à 2 points : $\xi(\boldsymbol{r}_1, \boldsymbol{r}_2) = 0$.

9.5.3 Flux de Poisson 2D.

Un flux de Poisson 2D est un flux ordinaire, homogène, uniforme et sans action à distance. Dans ces conditions la probabilité p_n de trouver n événements dans la surface ΔS est donnée par la loi de Poisson :

$$p_n(\Delta S) = \frac{(\lambda \Delta S)^n}{n!} e^{-\lambda \Delta S} \ . \tag{9.35}$$

La constante λ est la densité moyenne d'événements, autrement dit le nombre moyen d'événements par unité de surface.

Densité à n-points.

La densité à 1-point est le nombre moyen d'événements par unité de surface et donc $n_1(\boldsymbol{r}) = \lambda$ et elle ne dépend pas de $\boldsymbol{r}$. On montrerait de même que la densité à n-points est égale à λ^n, en particulier $n_2(\boldsymbol{r}_1, \boldsymbol{r}_2) = \lambda^2$ et la fonction de corrélation à 2 points $\xi(\boldsymbol{r}_1, \boldsymbol{r}_2)$ est donc nulle comme prévu.

Lois de la distance aux plus proches voisins.

Un point P de rayon vecteur $\boldsymbol{r}$ étant choisi, on cherche à connaître la densité de probabilité de la variable aléatoire D_1 qui est la distance de M au plus proche événement voisin. La probabilité pour qu'un disque de rayon d de surface ΔS ne contienne pas d'événement est donnée par la loi de Poisson :

$$\Pr\{d < D_1\} = \frac{(\lambda \Delta S)^0}{0!} e^{-\lambda \Delta S} = e^{-\lambda \Delta S} \ . \tag{9.36}$$

Sachant que $\Delta S = \pi d^2$ on en tire la fonction de répartition :

$$F_{D_1}(d) = \Pr\{D_1 \leq d\} = 1 - e^{-\lambda \pi d^2}, \tag{9.37}$$

et la densité de probabilité :

$$f_{D_1}(d) = 2\lambda \pi d \, e^{-\lambda \pi d^2} \ . \tag{9.38}$$

La figure 9.9 représente le graphe de cette densité.

Pour trouver la fonction de répartition de la distance D_n au n-ième plus proche voisin, il faut calculer la probabilité pour que le disque de rayon d contienne au moins n événements :

$$F_{D_n}(d) = \Pr\{D_n \leq d\} = \sum_{k=n}^{\infty} \frac{(\lambda \Delta S)^k}{k!} e^{-\lambda \Delta S}, \quad \Delta S = \pi d^2 \ . \tag{9.39}$$

On trouve la densité de probabilité par dérivation par rapport à d, soit :

$$f_{D_n}(d) = 2\lambda \pi d \frac{(\lambda \pi d^2)^{n-1}}{(n-1)!} e^{-\lambda \pi d^2} \ . \tag{9.40}$$

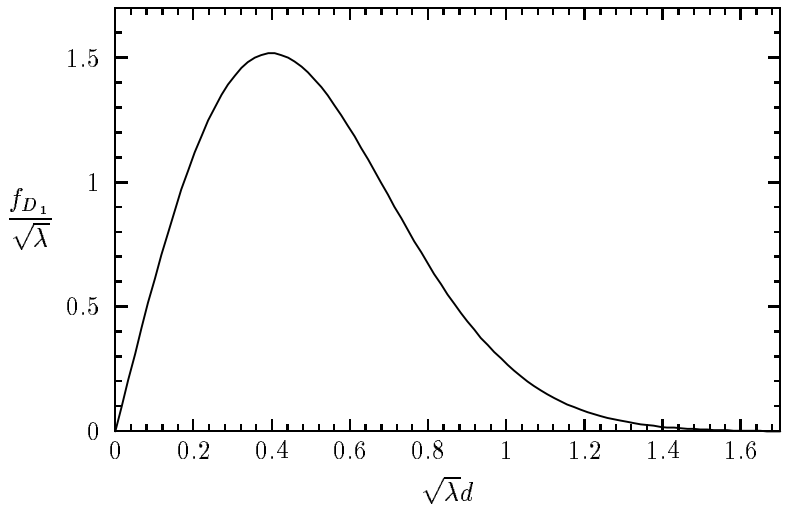


FIG. 9.9: *Densité de probabilité de la distance d'un point quelconque au plus proche événement voisin d'un flux de Poisson 2D de densité λ.*

Cette loi a pour moyenne et pour variance :

$$\mathrm{E}\{D_n\} = \frac{1}{\sqrt{\lambda\pi}} \frac{\Gamma(n+\frac{1}{2})}{(n-1)!}, \quad \mathrm{Var}(D_n) = \frac{1}{\lambda\pi}\left(n - \left[\frac{\Gamma(n+\frac{1}{2})}{(n-1)!}\right]^2\right). \qquad (9.41)$$

On a $\Gamma(n+\frac{1}{2}) = \frac{1}{2}\frac{3}{2}\cdots\frac{2n-1}{2}\sqrt{\pi}$, et on trouve en particulier pour la moyenne et la variance de la distance au plus proche voisin :

$$\mathrm{E}\{D_1\} = \frac{1}{2\sqrt{\lambda}}, \quad \mathrm{Var}(D_1) = \frac{1}{\lambda}\frac{4-\pi}{\pi}. \qquad (9.42)$$

9.6 Exercices et problèmes.

Exercice 9.1. *Lois conditionnelles.* Soit un flux de Poisson où les événements arrivent avec un taux constant λ. On s'intéresse aux cas où il y a exactement n événements dans l'intervalle de temps T.

Montrer que la loi gouvernant les dates d'arrivées ξ_k des événements dans cet intervalle de temps T, se déduit de la loi bêta. Plus spécifiquement montrer que :

$$\Pr\{\xi_k \le t | N = n\} = \frac{n!}{(k-1)!(n-k)!}\int_0^{t/T} u^{k-1}(1-u)^{n-k}du.$$

Exercice 9.2. On désire simuler un flux de Poisson de la façon suivante. On tire au hasard un nombre N suivant la loi de Poisson de moyenne $\mu = \lambda T$, on tire ensuite N nombres indépendants U_i au hasard suivant la loi uniforme entre 0 et T. Les valeurs λ et T sont données.

Si $U_{(1)}$ désigne le plus petit des U_i, montrer que cette variable aléatoire suit la loi exponentielle de paramètre λ dans l'intervalle $[0, T[$.

Dans l'éventualité où $N = 0$, on tire alors un nombre au hasard suivant la loi exponentielle de paramètre λ entre T et l'infini. On pose alors $U_{(1)}$ comme étant égal à ce nombre. Montrer qu'alors $U_{(1)}$ suit la loi exponentielle sur tout $\mathbb{R}^+$.

Exercice 9.3. *Cascade radioactive.* On considère un échantillon de matière radioactive entièrement constitué au temps $t = 0$ d'atomes d'espèce A_1. Les atomes A_1 se désintègrent en atomes d'espèce A_2 qui, à leur tour, se désintègrent en A_3 et ainsi de suite jusqu'à l'espèce A_{n+1} qui est stable. On suppose que le temps T_i de désintégration de A_i en A_{i+1}, $(i = 0, \ldots, n)$ suit une loi exponentielle de densité : $\lambda_k \exp\{-\lambda_k t\}$. Si T désigne le temps au bout duquel un atome A_1 devient stable, suite à n désintégrations successives indépendantes, montrer que la densité de probabilité $f_n(t)$ de cette variable aléatoire est égale à :

$$f_n(t) = \begin{cases} 0 & \text{si } t < 0\,, \\ (-1)^{n-1} \lambda_1 \lambda_2 \ldots \lambda_n \displaystyle\sum_{i=1}^{n} \frac{\exp\{-\lambda_i t\}}{\displaystyle\prod_{j \neq i} (\lambda_i - \lambda_j)} & \text{si } t \geq 0\,. \end{cases} \tag{9.43}$$

Deuxième partie

Statistique des variables aléatoires.

Chapitre 9

Echantillons et statistiques.

Nous abordons maintenant l'aspect statistique du traitement des données. Dans la première partie nous n'avons envisagé que le point de vue probabiliste : une loi est donnée et l'on cherche à caractériser une variable aléatoire ou une fonction de cette variable aléatoire, décrite par cette loi. Nous nous sommes efforcés, par exemple, de connaître les caractéristiques numériques d'une variable aléatoire en fonction des paramètres qui peuvent entrer dans l'expression de sa fonction de répartition. A présent, nous allons regarder les choses sous un autre angle ; on constate qu'il est dans la nature des choses que le résultat d'une expérience soit différent d'une fois à l'autre, et nous allons essayer d'appliquer la théorie des probabilités afin de rendre compte des fluctuations de ces résultats. Ce faisant on utilise la théorie des probabilités comme un modèle chargé de quantifier le caractère essentiellement aléatoire de la nature.

Clarifions ces notions en donnant un exemple. Un expérimentateur a observé une série de résultats $(x_1, \ldots, x_n)$ tous issus de la même expérience. Face à cette série de mesures, un statisticien pourra postuler que les x_i sont n issues indépendantes suivant, par exemple, une loi normale dont les paramètres μ et σ^2 sont inconnus. Il tentera, à partir des x_i, d'estimer ces paramètres. D'après la théorie des probabilités, on sait que les paramètres de la loi normale, σ^2 et μ, sont en fait une variance et une moyenne. Grâce à l'estimation de ces paramètres, l'expérimentateur aura une idée plus précise de la dispersion de ses données et de la valeur autour de laquelle elles se répartissent. Il pourrait également voir que, si ses erreurs de mesures étaient plus faibles ou ses mesures plus nombreuses, les valeurs x_i se concentreraient de plus en plus autour de μ. Si les choses se comportent bien ainsi, on peut raisonnablement dire que μ est en fait la valeur qu'il espérait mesurer et pour laquelle il dispose maintenant de l'estimation que lui a fournie le statisticien.

9.1 Les échantillons.

Le changement de point de vue que nous venons d'évoquer a pour conséquence que le statisticien utilise, pour désigner les mêmes objets, un vocabulaire différent de celui du probabiliste. Pour le statisticien, un ensemble $(X_1, \ldots, X_n)$ de n variables aléatoires est un « *échantillon* » et n en est sa « *taille* » ; il dit alors que c'est un échantillon de taille n ou encore un n-échantillon. Un ensemble

$(x_1, \ldots, x_n)$ de n nombres issus de ce n-échantillon est appelé une « *réalisation* » ou une « *observation* » de cet échantillon. Les lois de probabilité suivies par les variables aléatoires X_i sont appelées les « *populations parentes* ». Enfin, pour étayer ses prises de décisions, le statisticien utilise des fonctions $g(X_1, \ldots, X_n)$ des variables aléatoires X_i qu'il appelle des « *statistiques* ».

▶**Exemple 9.1.** *Problème typique exprimé avec le vocabulaire du statisticien.* Une expérience a conduit à observer n nombres $(x_1, \ldots, x_n)$, que l'on identifie à la réalisation d'un n-échantillon $(X_1, \ldots, X_n)$. On suppose que le n-échantillon est constitué de variables aléatoires indépendantes et issues de la même population parente de moyenne μ. Afin d'estimer μ on propose d'utiliser la statistique :

$$g(X_1, \ldots, X_n) = \frac{1}{n} \sum_{i=1}^{n} X_i \,. \tag{9.1}$$

Cette statistique, souvent notée $\overline{X}$, conduit à estimer la moyenne μ par la moyenne arithmétique des x_i :

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \,. \tag{9.2}$$

C'est un choix raisonnable puisque, sous réserve que la loi des grands nombres s'applique, $\overline{X}$ converge presque-sûrement vers μ lorsque la taille de l'échantillon tend vers l'infini.

Un n-échantillon $(X_1, \ldots, X_n)$ n'est donc rien de plus qu'un ensemble de n variables aléatoires. D'une façon générale, ces variables peuvent être indépendantes ou dépendantes, être issues de la même population ou de populations différentes. On peut aussi considérer les X_i comme les n composantes d'un vecteur aléatoire $\boldsymbol{X}$ suivant une loi à n dimensions. Dans ce dernier cas on a affaire à un échantillon de taille 1 issu d'une loi à n dimensions.

9.1.1 Les échantillons i.i.d.

L'échantillon pour lequel on possède le plus de résultats est l'échantillon « *indépendant et identiquement réparti,* » où les X_i qui le composent sont des variables aléatoires indépendantes les unes des autres et issues de la même population. On dira de manière abrégée qu'un tel échantillon est i.i.d d'après l'anglais : "independent and identically distributed". On identifiera une série de résultats $(x_1, \ldots, x_n)$ obtenus indépendamment les uns des autres, et dans les mêmes conditions expérimentales, avec la réalisation d'un échantillon i.i.d de taille n.

A partir de maintenant, sauf mention expresse du contraire, il ne sera question que d'échantillons i.i.d.

9.2 La fonction de vraisemblance.

En tant qu'ensemble de n variables aléatoires le n-échantillon possède, en général, une densité de probabilité : $f_n(x_1, \ldots, x_n; \theta_1, \ldots, \theta_k)$, où les valeurs $\theta_1, \ldots, \theta_k$ désignent les paramètres de la loi. D'un point de vue probabiliste, cette fonction f_n nous permet, les θ_j étant supposés connus, de calculer la densité de probabilité associée au point $(x_1, \ldots, x_n)$ de $\mathbb{R}^n$. L'approche statistique

inverse les rôles joués par les x_i et les θ_j elle suppose connues les valeurs $(x_1, \ldots,$-$x_n)$ et considère f_n comme fonction des θ_j. Cette densité, vue sous cet angle, reçoit le nom de « *fonction de vraisemblance* ». Afin de la distinguer de la densité, on la note L et on écrit $L(x_1, \ldots, x_n|\theta_1, \ldots, \theta_k)$ ou encore $L(\boldsymbol{x}|\boldsymbol{\theta})$ en suivant la convention habituelle qui veut qu'un symbole en caractère gras désigne un ensemble de valeurs.

La fonction de vraisemblance d'un n-échantillon i.i.d s'exprime simplement à partir de la densité $f(x|\boldsymbol{\theta})$ de la population parente. Les variables aléatoires X_i formant l'échantillon étant par hypothèse indépendantes, on a :

$$L(x_1, \ldots, x_n|\boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i; \boldsymbol{\theta}) \,. \tag{9.3}$$

▶**Exemple 9.2.** *Fonction de vraisemblance d'un échantillon de taille* 1. L'échantillon se réduit à une seule variable aléatoire X de densité $f(x;\theta)$. La fonction de vraisemblance $L(x|\theta)$ est alors égale à la densité de la population. Soit x_1 une réalisation de X ; sa fonction de vraisemblance vaut donc $L(x_1|\theta) = f(x_1;\theta)$. Si X suit, par exemple, une loi normale de moyenne μ et de variance unité, on aura :

$$L(x_1|\mu) = \frac{1}{\sqrt{2\pi}} \exp\{-\tfrac{1}{2}(x_1 - \mu)^2\} \,. \tag{9.4}$$

Le graphe de cette fonction est représenté sur la figure 9.1. Il faut noter en considérant cette figure que, pour la fonction de vraisemblance, μ est la variable et x_1 le paramètre.

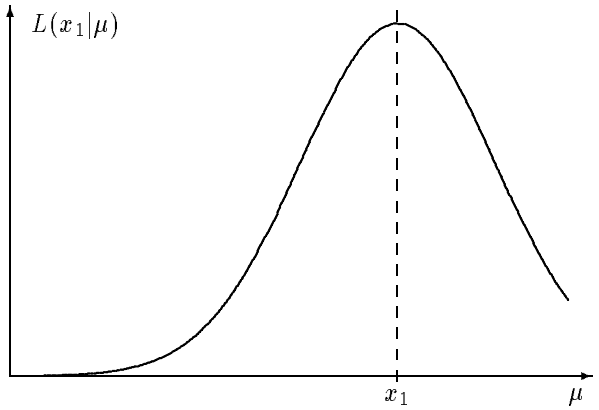


FIG. 9.1: *Fonction de vraisemblance d'un échantillon normal de taille 1. Dans ce cas la fonction de vraisemblance est égale à la densité de probabilité de la population parente. Sur cette figure la population parente suit la loi normale et x_1 désigne une réalisation de l'échantillon.*

▶**Exemple 9.3.** *Fonction de vraisemblance d'un échantillon issu de la loi exponentielle.* On suppose qu'un échantillon i.i.d de taille n est issu d'une loi exponentielle dont la densité de probabilité $f(x)$ dépend d'un paramètre inconnu θ. Plus précisément on a :

$$f(x) = \begin{cases} 0 \,, & \text{si } x < \theta \,; \\ \exp\{-(x-\theta)\} \,, & \text{si } x \geq \theta \,; \end{cases} \tag{9.5}$$

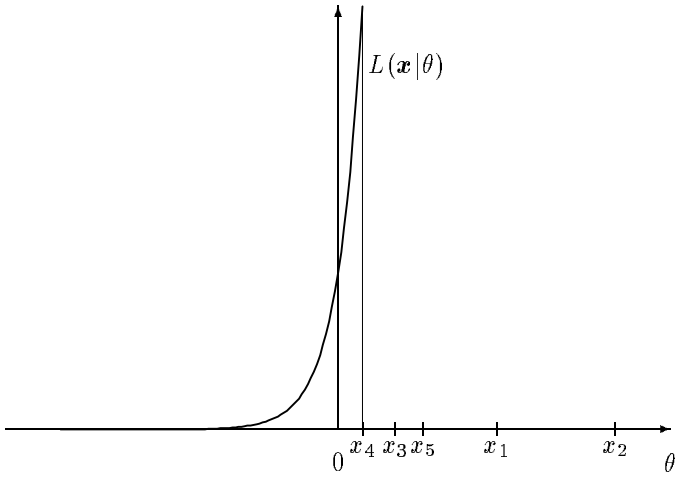


FIG. 9.2: *Fonction de vraisemblance d'un échantillon de taille 5 issu d'une loi exponentielle dont la densité de probabilité est donnée par l'équation* (9.5). *Le symbole* $\boldsymbol{x}$ *désigne l'ensemble des valeurs* $(x_1, \ldots, x_5)$, *le paramètre inconnu de cette loi:* θ *est inférieur ou égal à la plus petite de ces 5 valeurs, soit ici:* x_4. *Cette simulation a été faite avec* $\theta = 0$.

soit : $f(x|\theta) = \exp(\theta - x)\mathbf{1}_{\theta \leq x}$. Cette densité de probabilité est bornée à gauche par θ, aucune valeur de l'échantillon ne peut être inférieure à θ. Calculons la fonction de vraisemblance, il vient

$$\begin{aligned} L(x_1, \ldots, x_n|\theta) &= \prod_{i=1}^{n} \exp(\theta - x_i)\mathbf{1}_{\theta \leq x_i}\,, \\ &= \exp(n\theta - \textstyle\sum_{i=1}^{n} x_i) \displaystyle\prod_{i=1}^{n} \mathbf{1}_{\theta \leq x}\,, \\ &= \exp(n\theta - \textstyle\sum_{i=1}^{n} x_i)\mathbf{1}_{\min(x_1,\ldots,x_n)}\,, \end{aligned}$$

ou sous forme plus explicite :

$$L(x_1, \ldots, x_n|\theta) = \begin{cases} \exp\{n\theta - \sum_{i=1}^{n} x_i\}\,, & \text{si } \theta \leq \min(x_1, \ldots, x_n)\,; \\ 0\,, & \text{si } \theta > \min(x_1, \ldots, x_n)\,. \end{cases} \tag{9.6}$$

La fonction de vraisemblance L est bornée à droite par la plus petite valeur de l'échantillon. Ce comportement est conforme au fait que le paramètre θ ne peut pas être plus grand que la valeur minimum de l'échantillon : $\theta \leq \min(X_1, \ldots, X_n)$. La figure 9.3 représente une réalisation d'un échantillon de taille 5 issu d'une telle loi ainsi que la fonction de vraisemblance qui en découle.

Pour des raisons pratique on a avantage à considérer le logarithme de la fonction de vraisemblance plutôt que la fonction de vraisemblance elle-même. En effet dès que n est assez grand le produit (9.3) des densités de probabilité peut vite dépasser les possibilités de représentation d'un réel par un ordinateur. La fonction : $\ln L$ possède par ailleurs d'importants propriétés d'ordre théorique qui seront dégagées plus loin.

En tant que probabilité, la fonction L est positive ou nulle et dans le domaine où L est (presque-partout) strictement positive, l'application $L \to \ln L$ est

définie, continue, bijective et croissante. Considérer L ou $\ln L$ revient alors au même.

9.3 Les échantillons ordonnés.

On associe à un n-échantillon $(X_1, \ldots, X_n)$ un nouvel n-échantillon ordonné que l'on note $(X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)})$, et que l'on fabrique ainsi. On trie par ordre croissant les valeurs $(x_1, \ldots, x_n)$ des réalisations du n-échantillon. Soit $(x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)})$ le résultat de ce tri. On considère alors les $x_{(i)}$ comme les réalisations d'une certaine variable aléatoire $X_{(i)}$.

Il est clair que les variables ordonnées $X_{(i)}$ ne suivent pas nécessairement la même loi que les X_i et que de plus elles ne sont pas indépendantes, même si au départ les X_i l'étaient. Précisons maintenant la loi suivie par les $X_{(i)}$.

9.3.1 Loi suivie par les extrema d'un échantillon.

Considérons un échantillon i.i.d de taille n : $(X_1, \ldots, X_n)$, dont la population parente admet F pour fonction de répartition. Le minimum $X_{(1)}$ et le maximum $X_{(n)}$ de l'échantillon sont définis par :

$$X_{(1)} = \min(X_1, \ldots, X_n), \quad X_{(n)} = \max(X_1, \ldots, X_n). \tag{9.7}$$

Calculons les fonctions de répartition : $F_{(1)}$ et $F_{(n)}$ des variables aléatoires $X_{(1)}$ et $X_{(n)}$. Par définition de la fonction de répartition on a :

$$F_{(1)}(x) = \Pr\left\{X_{(1)} \leq x\right\}, \quad F_{(n)}(x) = \Pr\left\{X_{(n)} \leq x\right\}. \tag{9.8}$$

Nous commençons par la fonction $F_{(n)}$ qui est plus simple à évaluer. L'événement $\{X_{(n)} \leq x\}$, signifie que la plus grande des deux valeurs ne dépasse pas le seuil x. Pour que cette condition soit satisfaite, il faut et il suffit que *toutes* les variables X_i, $i = 1, \ldots, n$ ne dépassent pas le seuil x et les deux événements suivants sont alors équivalents :

$$\{X_{(n)} \leq x\} \iff \bigcap_{i=1}^{n} \{X_i \leq x\}. \tag{9.9}$$

Leurs probabilités associées sont par conséquent égales et comme par définition les événements $\{X_i \leq x\}$ sont indépendants il vient :

$$F_{(n)}(x) = \prod_{i=1}^{n} \Pr\{X_i \leq x\} = (F(x))^n. \tag{9.10}$$

Pour le calcul de $F_{(1)}$, il faut considérer l'événement $\{X_{(1)} \leq x\}$. Il signifie que la plus petite des valeurs de l'échantillon ne dépasse pas le seuil x, on ne sait rien sur les autres valeurs. Mais les deux événements : $\{X_{(1)} \leq x\}$ et $\{X_{(1)} > x\}$ forment un système complet d'événements incompatibles. On a alors $\Pr\left\{X_{(1)} \leq x\right\} + \Pr\left\{X_{(1)} > x\right\} = 1$ d'où on tire $1 - F_{(1)}(x) = \Pr\left\{X_{(1)} > x\right\}$. Pour que $X_{(1)}$ dépasse le seuil x il faut que *toutes* les variables X_i le dépassent aussi, il vient :

$$1 - F_{(1)}(x) = \Pr\left\{X_{(1)} > x\right\} = \prod_{i=1}^{n} \Pr\{X_i > x\} = (1 - F(x))^n, \tag{9.11}$$

Finalement les fonctions de répartition du minimum et du maximum de l'échantillon sont données par les expressions :

$$F_{(1)}(x) = 1 - (1 - F(x))^n \,, \tag{9.12a}$$
$$F_{(n)}(x) = (F(x))^n \,. \tag{9.12b}$$

▶**Exemple 9.4.** *Variables aléatoires indépendantes suivant les lois uniforme et exponentielle.* Illustrons les résultats précédents avec la loi uniforme sur $]0,1]$ pour laquelle on a $F(x) = x$ et donc :

$$F_{(1)}(x) = 1 - (1 - x)^n, \quad F_{(n)}(x) = x^n \,. \tag{9.13}$$

La figure 9.3 présente l'évolution de ces lois en fonction de la taille de l'échantillon.

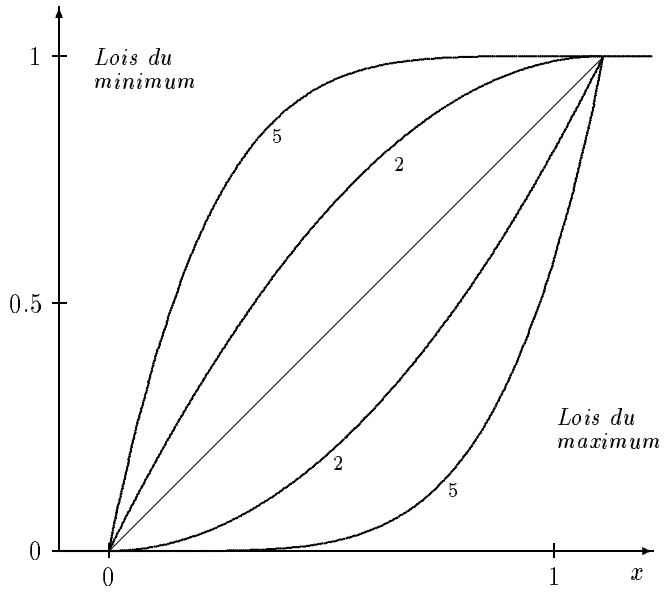


FIG. 9.3: *Fonction de répartition de la loi suivie par le minimum et le maximum d'un échantillon suivant la loi uniforme sur* $]0,1]$. *On a tracé ces fonctions pour un échantillon de taille 2 et pour un échantillon de taille 5. On a porté la taille de l'échantillon auprès de la fonction de répartition correspondante.*

Si les variables aléatoires suivent la loi exponentielle de fonction de répartition $F(x) = 1 - \exp(-\lambda x)$ pour $x \in [0, \infty[$, on a :

$$F_{(1)}(x) = 1 - [\exp(-\lambda x)]^n, \quad F_{(n)}(x) = [1 - \exp(-\lambda x)]^n \,. \tag{9.14}$$

9.3.2 Loi suivie par les variables ordonnées.

Posons nous maintenant le problème de déterminer la loi $F_{(k)}$ suivie par la k^{e} variable ordonnée $X_{(k)}$. Nous allons toujours nous placer dans le cas où l'échantillon $(X_1, \ldots, X_n)$ est i.i.d et suit une loi de fonction de répartition $F(x)$. Pour cela introduisons la variable aléatoire indicatrice $\mathbf{1}_{X_i \leq x}$ qui vaut 1 si $X_i \leq x$ et 0 si $X_i > x$. Cette variable suit une loi de Bernoulli de paramètre p qui est égal, par définition, à la probabilité de l'événement $\{X_i \leq x\}$, soit :

$p = \Pr\{X_k \leq x\}$. On a donc $p = F(x)$ et $\mathbf{1}_{X_i \leq x} = \mathcal{B}(1, F(x))$. Introduisons de plus la variable aléatoire ν qui vaut :

$$\nu = \sum_{i=1}^{n} \mathbf{1}_{X_i \leq x} \,. \tag{9.15}$$

Cette nouvelle variable compte le nombre de variables X_i qui n'ont pas dépassé le seuil x. Les variables $\mathbf{1}_{X_i \leq x}$ sont indépendantes car elle ne sont fonctions que des variables X_i qui sont elles-mêmes indépendantes. En tant que somme de n variables aléatoires de Bernoulli indépendantes et de même paramètre p, ν suit une loi binomiale d'expression :

$$\Pr\{\nu = r\} = C_n^r \, p^r (1-p)^{n-r} \tag{9.16}$$

Pour que l'événement $\{X_{(k)} \leq x\}$ soit réalisé, il faut que l'on ait au moins k variables X_i qui soient inférieures ou égales à x. En d'autres termes, il faut que le nombre ν de variables qui n'ont pas dépassé x soit supérieur ou égal à k. Cela correspond à l'événement $\{\nu \geq k\}$, qui a pour probabilité :

$$\Pr\{\nu \geq k\} = \sum_{r=k}^{n} \Pr\{\nu = r\} \,, \tag{9.17}$$

ce qui nous permet, à l'aide de l'équation (9.16), de trouver l'expression cherchée :

$$F_{(k)}(x) = \sum_{r=k}^{n} C_n^r \, F^r(x)(1 - F(x))^{n-r} \,. \tag{9.18}$$

L'étude des fonctions eulériennes nous permet d'écrire ce résultat sous une autre forme. En effet on y apprend que :

$$\sum_{r=k}^{n} C_n^r p^r (1-p)^{n-r} = \frac{n!}{(k-1)!\,(n-k)!} \int_0^p u^{k-1}(1-u)^{n-k} du \,. \tag{9.19}$$

Le second membre de cette égalité s'exprime à l'aide de la fonction bêta incomplète normalisée (voir appendice A) qui par définition est égale à :

$$I_x(a,b) = \frac{1}{B(a,b)} \int_0^x u^{a-1}(1-u)^{b-1} du \,. \tag{9.20}$$

Nous pouvons alors exprimer la fonction de répartition de $X_{(k)}$ sous la forme :

$$F_{(k)}(x) = I_{F(x)}(k, n-k+1) \,. \tag{9.21}$$

Comme le montre l'équation (9.18), $F_{(k)}$ est un polynôme en $F(x)$ de degré n. Pour calculer l'expression de $F_{(k)}$ il suffit d'avoir à sa disposition une table ou un programme permettant de calculer la fonction bêta incomplète normalisée. La valeur de x étant connue, on calcule $u = F(x)$ et l'on cherche I_u dans la table. La table 9.1 est extraite d'une table de $I_u(a,b)$ pour $a = 15$ et $b = 10$.

▶**Exemple 9.5.** *Variables uniformes ordonnées.* Si les variables aléatoires ($U_1, \ldots,$ U_n) suivent la loi uniforme de fonction de répartition $F(u) = u$ sur $[0,1]$, les variables

u	$I_u(15,10)$	u	$I_u(15,10)$	u	$I_u(15,10)$
0.41	0.0275200	0.51	0.1782786	0.61	0.5294736
0.42	0.0345998	0.52	0.2051618	0.62	0.5698155
0.43	0.0430593	0.53	0.2343327	0.63	0.6097143
0.44	0.0530619	0.54	0.2657005	0.64	0.6487830
0.45	0.0647690	0.55	0.2991267	0.65	0.686650
0.46	0.0783350	0.56	0.3344253	0.66	0.722968
0.47	0.0939020	0.57	0.3713639	0.67	0.757427
0.48	0.1115943	0.58	0.4096672	0.68	0.789758
0.49	0.1315124	0.59	0.4490213	0.69	0.819743
0.50	0.1537281	0.60	0.4890802	0.70	0.847218

TAB. 9.1: *Extrait d'une table de la fonction bêta incomplète $I_u(a,b)$ pour des valeurs de u comprises entre* 0.41 *et* 0.70*, pour a=15 et b=10. Voir "Tables of the incomplete beta-function" éditées par K. Pearson [57]. La présente table pourrait servir à déterminer la loi suivie par la $a = 15^e$ variable ordonnée $\xi_{(15)}$ d'un échantillon de taille $n = a + b - 1 = 24$.*

ordonnées $U_{(k)}$ suivent, d'après (9.21), une loi de fonction de répartition $I_u(k, n-k+1)$. Les variables ordonnées suivent par conséquent une loi bêta :

$$U_{(k)} = \beta(k, n-k+1)\,. \tag{9.22}$$

Si la variable aléatoire X est continue de fonction de répartition F, on peut toujours lui associer la variable aléatoire uniforme $U = F(X)$. Cette remarque nous permet d'exprimer les quantiles $x_{\alpha,(k)}$ des variables ordonnées $X_{(k)}$ d'après les quantiles de la loi bêta correspondante. L'ordre étant conservé par l'application F, on a : $F(X_{\alpha,(k)}) = u_{\alpha,(k)}$, où $u_{\alpha,(k)}$ désigne le quantile d'ordre α de la loi $\beta(k, n-k+1)$.

Soit, par exemple, un échantillon $(X_1, \ldots, X_n)$ de taille $n = 24$ et de population parente exponentielle : $F(x) = 1 - \exp(-\lambda x)$. On désire calculer la médiane $x_{0.5,(15)}$ de la 15^e variable ordonnée. Par interpolation dans la table 9.1 on trouve $u_{0.5,(15)} \approx 0.6027$, d'où $\lambda x_{0.5,(15)} = -\ln(1 - u_{0.5,(15)}) \approx 0.9231$. Il y a donc une chance sur deux pour que la 15^e parmi 24 variables exponentielles dépasse le seuil $\approx 0.9231\lambda^{-1}$.

▶**Exemple 9.6.** *Apparition de pannes sur des composants fonctionnant en batterie.* Une expérience dépend du bon fonctionnement d'un composant électronique sujet à des pannes. La durée de vie moyenne τ de ce composant se trouve être beaucoup trop courte par rapport au temps que doit durer l'expérience. Afin d'assurer le bon déroulement de l'expérience, on installe 5 composants identiques, fonctionnant en parallèle. Pour que l'expérience fonctionne sans interruption, on décide qu'au bout du temps T où le 4^e composant tombe en panne, on remplace l'ensemble des composants. On demande la loi suivie par le temps T au bout duquel il faut remplacer les composants.

La loi qui préside à l'apparition des pannes est, sous des hypothèses assez générales, la loi exponentielle, de fonction de répartition :

$$F(t) = 1 - e^{-\lambda t}\,. \tag{9.23}$$

Dans cette expression λ^{-1} est la valeur moyenne de la loi, c'est-à-dire la durée de vie moyenne d'un composant. Il vient donc $\tau = \lambda^{-1}$. Soit T_i la variable aléatoire représentant le temps d'apparition d'une panne sur le composant numéro i, et $T_{(i)}$ les variables aléatoires T_i triées par ordre croissant. Avec ces définitions, T, temps de

remplacement de l'ensemble des composants, est égal à $T_{(4)}$. La loi suivie par T a donc pour fonction de répartition $\Pr\{T \leq t\} = F_{(4)}(t)$ qui est donnée par l'expression :

$$F_{(4)}(t) = I_{F(t)}(4, 2)\,, \tag{9.24}$$

dont la forme analytique est donnée par l'équation (9.20) et vaut :

$$F_{(4)}(t) = \frac{5!}{(4-1)!\,(5-4)!} \int_0^{1-e^{t/\tau}} u^3(1-u)\,du \tag{9.25}$$

$$= 5(1-e^{-t/\tau})^4 - 4(1-e^{-t/\tau})^5 \tag{9.26}$$

$$= 1 - 10e^{-2t/\tau} + 20e^{-3t/\tau} - 15e^{-4t/\tau} + 4e^{-5t/\tau}\,. \tag{9.27}$$

Le graphe de cette fonction est donné par la figure 9.4.

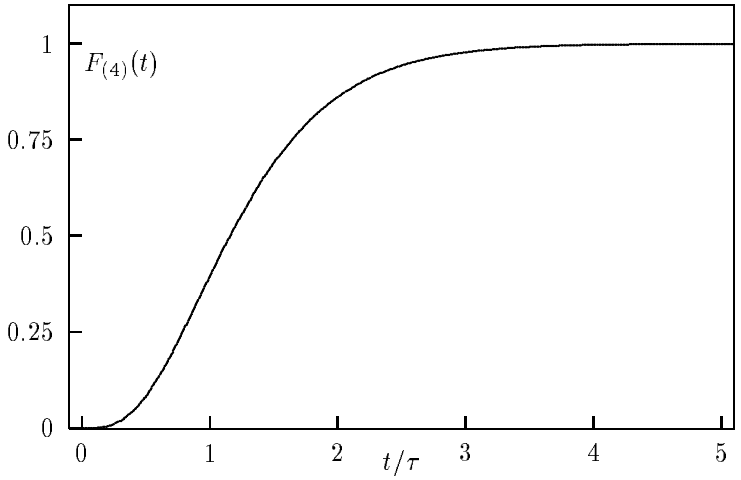


FIG. 9.4: *Fonction de répartition de la loi suivie par le temps au bout duquel il faut remplacer une batterie de 5 composants fonctionnant en parallèle, quand on décide de les remplacer lorsque le 4^e composant subit une panne. Sur ce graphique, τ est la durée de vie moyenne d'un composant.*

9.4 La fonction de répartition empirique.

La fonction de répartition empirique F_n est une approximation de la fonction de répartition F de la population parente basée sur l'échantillon ($X_1, \ldots,$-X_n). Nous donnons ci-dessous la définition de cette fonction réservant pour plus tard (chapitre 16) la question d'apprécier quantitativement la qualité de cette approximation.

9.4.1 Une définition « naturelle » de F_n.

Nous avons vu (équation (5.38), page 72) que la fonction de répartition pouvait s'interpréter comme espérance mathématique de l'indicatrice des variables aléatoires X_i. On a en effet :

$$F(x) = \mathrm{E}\,\{\mathbf{1}_{X \leq x}\} = \int_{-\infty}^{x} dF\,. \tag{9.28}$$

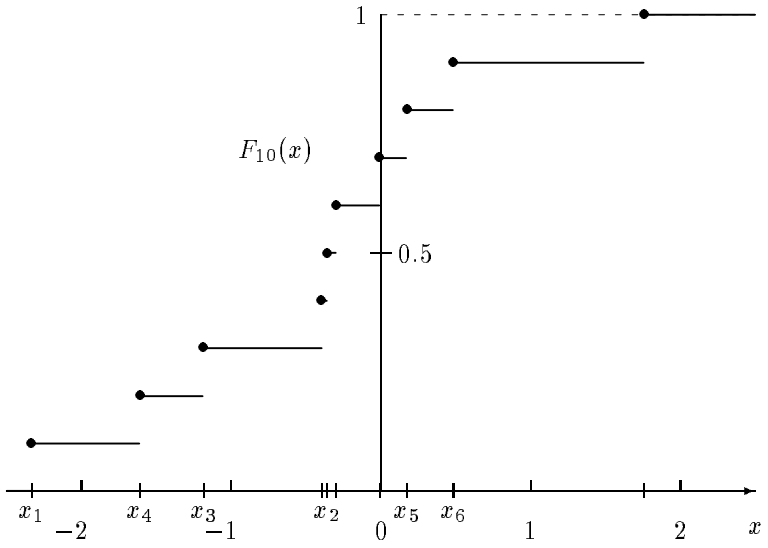


FIG. 9.5: *Exemple d'une réalisation de la fonction de répartition empirique $F_n(x)$ d'un échantillon normal réduit de taille $n = 10$. Pour plus de clarté seuls les 6 premières valeurs de l'échantillon ont été identifiées sur l'axe des x.*

En remplaçant l'opérateur espérance mathématique E{} par la moyenne arithmétique $\frac{1}{n}\sum_{i=1}^{n}$, on obtient une approximation dite « *naturelle* » de la fonction de répartition, soit :

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{X_i \leq x} = \int_{-\infty}^{x} dF_n \,. \tag{9.29}$$

La quantité $F_n(x)$ est la proportion de variables qui ne dépassent pas la valeur x. On appelle « *fonction de répartition empirique* » la fonction F_n ainsi définie, on lui donne parfois aussi le nom de « *courbe cumulative* ». Cette fonction est la somme de fonctions « en escaliers » présentant un saut d'amplitude $\frac{1}{n}$ pour chaque valeur X_i de l'échantillon (voir figure 9.5).

Pour chaque valeur de x fixée, $F_n(x)$ est une variable aléatoire que nous allons maintenant étudier.

9.4.2 Loi suivie par la variable aléatoire $F_n(x)$.

La variable aléatoire $F_n(x)$ est une variable aléatoire discrète, ne pouvant prendre que les valeurs $0, \frac{1}{n}, \ldots, \frac{k}{n}, \ldots, 1$. La variable aléatoire $nF_n(x)$ est également une variable discrète à valeurs dans $\{0, 1, \ldots, n\}$, elle est la somme des n variables $\mathbf{1}_{X_i \leq x}$. Ces variables aléatoires indicatrices sont par définition des variables de Bernoulli et nous avons vu plus haut que leur paramètre p était égal à $F(x)$. Elles sont de plus i.i.d car les variables X_i sont elles-mêmes i.i.d. La variable $nF_n(x)$ est alors la somme de n variables indépendantes de Bernoulli de même paramètre p et suit, par conséquent, la loi binomiale ($nF_n(x)$ est identique à la variable ν de l'équation (9.15) ci-dessus) :

$$nF_n(x) = \mathcal{B}(n, F(x)) \,. \tag{9.30}$$

D'où on tire pour la variable $F_n(x)$:

$$\Pr\Big\{F_n(x) = \frac{k}{n}\Big\} = C_n^k (F(x))^k (1-F(x))^{n-k} \,. \tag{9.31}$$

La variable aléatoire $F_n(x) = \frac{1}{n}\mathcal{B}(n, F(x))$ possède la moyenne $F(x)$ et la variance $\frac{1}{n}F(x)(1-F(x))$, en effet :

$$\mathrm{E}\{F_n(x)\} = \frac{1}{n}\,\mathrm{E}\,\{nF_n(x)\} = \frac{1}{n}nF(x) = F(x)\,, \tag{9.32}$$

$$\mathrm{Var}(F_n(x)) = \frac{1}{n^2}\,\mathrm{Var}(nF_n(x)) = \frac{1}{n^2}nF(x)(1-F(x)) = \frac{1}{n}F(x)(1-F(x))\,. \tag{9.33}$$

9.4.3 Convergence de F_n vers F.

D'après la loi forte des grands nombres (voir théorème 7.10, page 118), la variable aléatoire $F_n(x)$ en tant que somme de n variables aléatoires i.i.d converge presque-sûrement vers sa moyenne lorsque $n \to \infty$. Les calculs précédents ont montré que cette moyenne existe pour tout x et vaut $F(x)$, on a donc :

$$\forall x, \quad F_n(x) \xrightarrow{\text{p.s.}} F(x)\,. \tag{9.34}$$

La convergence presque-sûre de $F_n(x)$ vers $F(x)$ en tout x n'assure naturellement pas la convergence uniforme (presque-sûre) de la fonction F_n vers F, cependant V. I. Glivenko et F. P. Cantelli ont pu démontrer le théorème suivant :

Théorème 9.1. *(Glivenko-Cantelli, 1933.) La fonction de répartition empirique F_n issue d'un échantillon i.i.d $(X_1, \ldots, X_n)$ de fonction de répartition F, converge presque-sûrement vers F de façon uniforme en x, lorsque $n \to \infty$. C'est-à-dire :*

$$\sup_x |F_n(x) - F(x)| \xrightarrow{p.s.} 0\,. \tag{9.35}$$

On trouvera la démonstration de ce théorème au chapitre 7 §8 de l'ouvrage de Rényi [62].

La version du théorème central limite concernant la convergence en loi d'une variable binomiale vers la loi normale (voir théorème 7.19, page 123) nous permet de décrire comment $F_n(x)$ tend vers $F(x)$ lorsque $n \to \infty$:

$$\sqrt{n}\frac{F_n(x) - F(x)}{\sqrt{F(x)(1-F(x))}} \xrightarrow{\text{loi}} \mathcal{N}(0,1)\,. \tag{9.36}$$

Ce qui signifie que les écarts à la courbe théorique tendent à se comporter comme des variables aléatoires normales. La figure 9.4.3 représente une réalisation de l'écart réduit entre $F_n(x)$ et $F(x)$ pour un échantillon de taille 100 extrait d'une population parente uniforme.

9.5 Statistiques associées à un échantillon.

Comme nous l'avons déjà mentionné, une statistique est une fonction des variables aléatoires composant le n-échantillon. Une statistique est donc elle-même une variable aléatoire et obéit à ce titre une certaine loi. Si elle possède

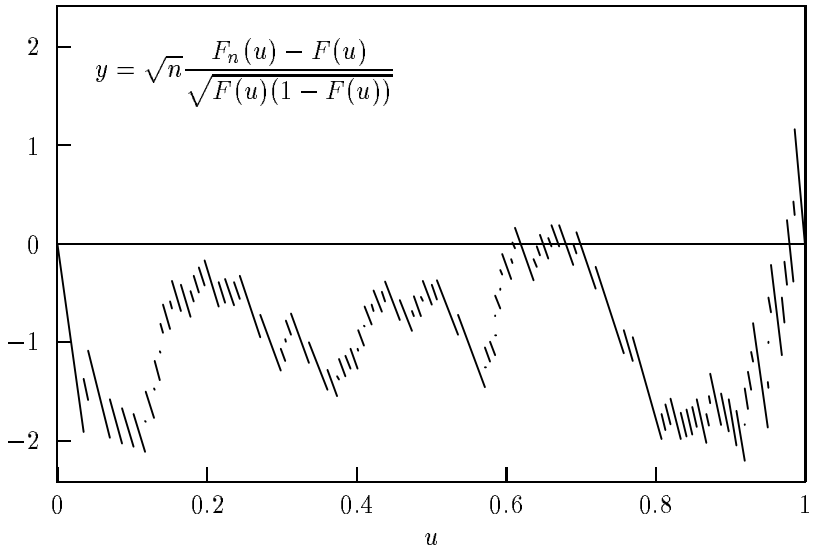


FIG. 9.6: *Ecarts réduits entre la fonction de répartition empirique $F_n(u)$ et la fonction de répartition d'un variable aléatoire uniforme $F(u) = u$. La fonction de répartition empirique a été calculée à l'aide d'un échantillon de taille 100. Pour x donné, les écarts réduits tendent en loi vers une loi normale : $\mathcal{N}(0,1)$. La fonction aléatoire dont une réalisation est représentée ici est ce que l'on apelle « un pont brownien ».*

une densité de probabilité cette dernière peut, en principe, être calculée d'après les règles exposées au chapitre concernant le changement de variable aléatoire.

On distingue habituellement deux types de statistiques : 1) celles qui sont fonction des variables X_i de l'échantillon, 2) celles qu'on appelle « *statistiques d'ordre* », et qui sont fonction des variables ordonnées $X_{(i)}$. Nous donnons dans les sections suivantes les statistiques les plus souvent utilisées, mais auparavant nous voulons dégager le lien profond qui existe entre les statistiques et la fonction de répartition empirique.

9.5.1 Les statistiques en tant que fonctionnelles.

Une statistique T est une fonction de l'échantillon, nous la notons $T = T(X_1, \ldots, X_n)$, et nous n'envisageons, en règle générale, que des statistiques fonctions d'échantillons i.i.d.

Considérons à titre d'exemple deux statistiques usuelles : « la moyenne empirique » M et la « médiane empirique » $Q_{0.5}$. Ces statistiques sont définies ainsi :

$$M = \frac{1}{n}\sum_{i=1}^{n} X_i\,, \quad Q_{0.5} = \begin{cases} X_{(p+1)}\,, & \text{si } n = 2p+1\,; \\ \frac{1}{2}(X_{(p+1)} + X_{(p)})\,, & \text{si } n = 2p\,. \end{cases}$$

D'après la définition de la fonction de répartition empirique F_n et les propriétés de l'intégrale de Stieltjes on peut redéfinir M en tant que fonctionnelle de F_n.

Cette fonctionnelle s'exprime explicitement à l'aide d'une intégrale :

$$M = \int x \, dF_n \, . \tag{9.37}$$

La statistique $Q_{0.5}$ ne s'exprime pas à l'aide d'une intégrale mais elle n'en est pas moins une fonctionnelle de F_n, on peut écrire en effet :

$$Q_{0.5} = \begin{cases} \min\{Q \mid F_n(Q) = \frac{1}{2} + \frac{1}{n}\} \, , & \text{si } n \text{ est pair} \, ; \\ \frac{1}{2}(Q_a + Q_b) \, , & \\ Q_a = \min\{Q \mid F_n(Q) = \frac{1}{2} - \frac{1}{2n}\} \, , & \\ Q_b = \min\{Q \mid F_n(Q) = \frac{1}{2} + \frac{1}{2n}\} \, , & \text{si } n \text{ est impair} \, . \end{cases} \tag{9.38}$$

Une fonctionnelle est une application qui à une fonction fait correspondre un nombre, soit A cette application, $A : F \mapsto x \in \mathbb{R}$, où F appartient à l'ensemble des fonctions qui sont des fonctions de répartition. Afin de pouvoir employer les outils de l'analyse il faut définir des distances dans les espaces de départ et d'arrivée. L'espace de départ est celui des réels et la distance habituelle est $d(x, y) = |x - y|$. Dans l'espace des fonctions il existe plusieurs définitions de la distance, nous considérons ici la distance de Kolmogorov :

$$d_K(F, G) = \sup_x |F(x) - G(x)| \, . \tag{9.39}$$

Les statistiques M et $Q_{0.5}$ appartiennent, comme l'immense majorité des statistiques usuelles, a seulement deux classes suivant la nature de la fonctionnelle qui les définissent.

- **Les statistiques de classe I.** Ce sont les statistiques $T = A(F_n)$ qui dépendent explicitement d'une intégrale, soit :

  $$T = h\Big(\int g(x) \, dF_n(x) \Big) \, ,$$

  où g est une fonction mesurable-Borel et h une fonction continue au voisinage du nombre $\int g(x) \, dF_n(x)$. Grossièrement une fonction est mesurable-Borel si elle est intégrable (une fonction continue est mesurable-Borel).

- **Les statistiques de classe II.** Ce sont des statistiques définies à l'aide d'une fonctionnelle continue au voisinage du « point » F_n. Le voisinage étant défini, pour les fonctions de répartition, au sens de la distance de Kolmogorov.

9.5.2 Convergence des statistiques.

Il est clair que la moyenne empirique et la médiane empirique ne seront des statistiques intéressantes, d'un point de vue pratique, que si elles convergent respectivement vers la moyenne et vers la médiane de la population parente.

Ce fait est établi facilement pour M à l'aide d'une version quelconque de la loi des grands nombres. La moyenne empirique M est en effet une variable aléatoire somme de n variables aléatoires indépendantes, et si ces dernières possèdent la même moyenne μ alors on sait, d'après la version forte due à Kolmogorov de la

loi des grands nombres (théorème 7.17, page 120), que M converge presque-sûrement vers μ.

Pour la médiane empirique la convergence est une simple conséquence du théorème de Glivenko-Cantelli lorsque les variables aléatoires de l'échantillon sont continues. Ne considérons pour simplifier que des échantillons de taille paire, si $x_{0.5}$ désigne la médiane de la population, on a :

$$\sup_x |F_n(x) - F(x)| \geq |F_n(Q_{0.5}) - F(Q_{0.5})| \xrightarrow{\text{p.s.}} 0\,,$$
$$|F_n(Q_{0.5}) - F(Q_{0.5})| = |\tfrac{1}{2} + \tfrac{1}{n} - F(Q_{0.5})| \xrightarrow{\text{p.s.}} 0\,,$$
$$F(Q_{0.5}) \xrightarrow{\text{p.s.}} \tfrac{1}{2}\,,\ Q_{0.5} \xrightarrow{\text{p.s.}} F^{-1}(\tfrac{1}{2}) = x_{0.5}\,.$$

Les démonstrations précédentes sont deux cas particuliers d'un théorème plus général :

Théorème 9.2. *Soit une statistique $T = T(X_1, \ldots, X_n)$ calculée à partir d'un échantillon i.i.d de taille n et de population parente admettant F comme fonction de répartition. Si T est une fonctionnelle de la fonction de répartition empirique : $T = A(F_n)$, et si cette fonctionnelle appartient à au moins une des deux classes précédemment définies, alors T converge presque-sûrement vers $A(F)$ (si cette dernière quantité existe) :*

$$A(F) < \infty \Rightarrow A(F_n) \xrightarrow{p.s.} A(F)\,. \tag{9.40}$$

On trouvera la démonstration de ce théorème au chapitre 1 §3 de l'ouvrage de A. Borovkov [11]. Nous définissons maintenant les moments empiriques de l'échantillon.

9.6 Moments de l'échantillon.

On appelle « *moments de l'échantillon* » ou « *moments échantillonnés* » ou encore « *moments empiriques* », les statistiques M'_k définies comme moyenne arithmétique des puissances k des X_i :

$$M'_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k\,, \quad k \geq 1. \tag{9.41}$$

On donne au moment M'_1 le nom de moyenne de l'échantillon (ou de moyenne empirique) et on le note M. On introduit également les moments centrés M_k de l'échantillon, qui sont des statistiques définies par :

$$M_k = \frac{1}{n}\sum_{i=1}^{n}(X_i - M)^k\,, \quad k \geq 2\,. \tag{9.42}$$

Parmi les moments centrés, on distingue M_2 la variance de l'échantillon que l'on note S'^2 et qui vaut donc :

$$S'^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - M)^2\,. \tag{9.43}$$

On introduit, de plus, la variance empirique modifiée S^2, qui est d'une plus grande importance pratique que S'^2, et qui a pour expression :

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - M)^2 \,. \tag{9.44}$$

La moyenne M d'un échantillon $(X_1, \ldots, X_n)$ est souvent notée $\overline{X}$ de façon à pouvoir la distinguer de la moyenne d'un échantillon différent $(Y_1, \ldots, Y_n)$ dont la moyenne empirique sera alors notée $\overline{Y}$. La taille de l'échantillon est, si nécessaire, portée en indice de la statistique considérée, par exemple $\overline{X}_n$ pour la moyenne empirique ou S_n^2 pour la variance empirique modifiée. On note les réalisations d'une statistique, par la lettre minuscule qui lui correspond, ainsi $\overline{x}$ désigne une réalisation de la moyenne $\overline{X}$. Soit :

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \,. \tag{9.45}$$

Un problème de notation. En dépit des règles énoncées ci-dessus, il faut bien reconnaître qu'il règne, dans la littérature, une certaine confusion d'écriture et de vocabulaire. Bien souvent on note de la même façon et on parle de la même manière d'une variable aléatoire et de sa réalisation. Ainsi on dit souvent « échantillon » quand il s'agit en fait d'une de ses réalisations, ou encore on dit « moyenne empirique » à propos du scalaire $\overline{x}$ alors qu'on devrait dire « réalisation de la moyenne empirique $\overline{X}$ ». Une telle précision dans le vocabulaire est quelque peu lourde et l'on est souvent tenté de commettre ce genre d'abus de langage. Bien que le contexte dissipe, en général, l'équivoque il est bon de faire clairement la distinction dans son esprit à défaut de toujours la trouver dans le texte.

9.6.1 Convergence des moments empiriques.

Les moments M'_k sont de toute évidence des statistiques de classe I, pour les moments centrés M_k il suffit de remarquer qu'ils s'écrivent : $M^k = \int (x - M)^k \, dF_n(x) = \int g(x)\, dF_n(x)$ avec $g(x) = \left(x - \int x \, dF_n(x)\right)^k$. On a donc :

$$\mathrm{E}\{|X|^k\} < \infty \Rightarrow \begin{cases} M'_k \xrightarrow{\text{p.s.}} \mu'_k \,, \\ M_k \xrightarrow{\text{p.s.}} \mu_k \,, \end{cases} \tag{9.46}$$

lorsque la taille de l'échantillon tend vers l'infini. En particulier $S'^2 \xrightarrow{\text{p.s.}} \sigma^2$ ainsi que S^2. Les fonctions des momemts empiriques convergent aussi vers leur équivalents de la population (s'ils existent). On a, par exemple, pour les coefficients d'asymétrie et d'aplatissement empiriques :

$$\frac{M_3}{M_2^{3/2}} \xrightarrow{\text{p.s.}} \gamma_1 \,, \quad \frac{M_4}{M_2^2} - 3 \xrightarrow{\text{p.s.}} \gamma_2 \,, \tag{9.47}$$

où γ_1 et γ_2 désignent les coefficients d'asymétrie et d'aplatissement de la population parente.

Le théorème central limite s'applique si les variables aléatoires X^k possèdent une variance. On a : $\mathrm{Var}(X^k) = \mathrm{E}\{(X^k - \mathrm{E}\{X^k\})^2\} = \mu'_{2k} - \mu'^2_k$, alors la variance

des X^k existe si la population parente possède des moments jusqu'à l'ordre $2k$, dans ces conditions on a :

$$\mathrm{E}\{|X|^{2k}\} < \infty \Rightarrow \sqrt{n}\frac{M'_k - \mu'_k}{\mathrm{Var}(X^k)^{\frac{1}{2}}} = \frac{M'_k - \mu'_k}{\mathrm{Var}(M'_k)^{\frac{1}{2}}} \xrightarrow{\text{loi}} \mathcal{N}(0,1)\,. \tag{9.48}$$

On a également :

$$\mathrm{E}\{|X|^{2k}\} < \infty \Rightarrow \frac{M_k - \mu_k}{\mathrm{Var}(M_k)^{\frac{1}{2}}} \xrightarrow{\text{loi}} \mathcal{N}(0,1)\,. \tag{9.49}$$

9.6.2 Caractéristiques numériques des moments empiriques.

Espérance et variance de la moyenne empirique. Nous avons déjà trouvé ces quantités à titre d'exemple (voir page 77). Rappelons ici les résultats. L'opérateur espérance mathématique étant linéaire on a immédiatement, à condition que la moyenne μ de la population existe :

$$\mathrm{E}\{M\} = \mu\,. \tag{9.50}$$

Pour la variance de la moyenne empirique on a (sous réserve d'existence de la variance σ^2 de la population) :

$$\mathrm{Var}(M) = \frac{\sigma^2}{n}\,. \tag{9.51}$$

Espérance et variance de la variance empirique. On démontre que l'espérance et la variance de S'^2 sont données par les expressions :

$$\mathrm{E}\{S'^2\} = (1 - \frac{1}{n})\mu_2\,, \tag{9.52}$$

$$\mathrm{Var}(S'^2) = \frac{\mu_4 - \mu_2^2}{n} - \frac{2(\mu_4 - 2\mu_2^2)}{n^2} + \frac{\mu_4 - 3\mu_2^2}{n^3}\,. \tag{9.53}$$

Espérance variance et covariance des moments empiriques.

On démontre également les résultats suivants (voir Kendall & Stuart [40] chap. 10.4 et 10.5). Sous réserve d'existence des moments de la population parente apparaissant dans les expressions ci-dessous, les espérances, les variances et les covariances des moments empiriques sont données par les formules :

$$\mathrm{E}\{M'_r\} = \mu'_r\,, \tag{9.54}$$

$$\mathrm{Var}(M'_r) = \frac{1}{n}(\mu'_{2r} - \mu'^2_r)\,, \tag{9.55}$$

$$\mathrm{Cov}(M'_q, M'_r) = \frac{1}{n}(\mu'_{q+r} - \mu'_q\mu'_r)\,, \tag{9.56}$$

$$\mathrm{E}\{M_r\} = \mu_r + O(n^{-\frac{1}{2}})\,, \tag{9.57}$$

$$\mathrm{Var}(M_r) = \frac{1}{n}(\mu_{2r} - \mu_r^2 + r^2\mu_2^2\mu_{r-1}^2 - 2r\mu_{r-1}\mu_{r+1}) + O(n^{-2})\,, \tag{9.58}$$

$$\mathrm{Cov}(M_q, M_r) = \frac{1}{n}(\mu_{r+q} - \mu_r\mu_q + rq\mu_2\mu_{r-1}\mu_{q-1} - r\mu_{r-1}\mu_{q+1} - q\mu_{r+1}\mu_{q-1}) + O(n^{-2}) \tag{9.59}$$

Il faut remarquer que les variances de M_r et de M'_r ne sont finies que lorsque la loi possède des moments jusqu'à l'ordre $2r$.

9.7 Statistiques d'ordre.

Une statistique d'ordre est une statistique calculée à partir de l'échantillon ordonné, c'est donc une fonction des variables $X_{(i)}$. La variable $X_{(i)}$ elle-même est une statistique d'ordre mais il en existe bien d'autres, nous en donnons et rappelons quelques unes ci-dessous.

- Le **rang** R_m est une statistique d'ordre égale à la place qu'occupe X_m dans l'échantillon ordonné. Si, par exemple, on a observé : $(x_4 < x_3 < x_5 < x_1 < x_2)$, le rang de X_3 est alors égal à 2 pour cette réalisation $(r_3 = 2)$.

- La **fréquence empirique** $N_n(x)$ est égale au nombre de variables de l'échantillon qui sont inférieures ou égales à un x donné. D'après la définition de la fonction de répartition empirique ce nombre est égal à $nF_n(x)$, c'est une variable binomiale de paramètre $p = F(x)$:

$$N_n(x) = \mathcal{B}(n, F(x)) . \tag{9.60}$$

- Les **valeurs extrêmes**. Les variables aléatoires $X_{(1)}$ et $X_{(n)}$ issues d'un échantillon de taille n sont des statistiques dites des extrêmes. Il s'agit ici respectivement du minimum et du maximum de l'échantillon :

$$X_{(1)} = \min(X_1, \ldots, X_n), \quad X_{(n)} = \max(X_1, \ldots, X_n) . \tag{9.61}$$

- Le **point milieu** est le point à égale distance des valeurs extrêmes :

$$P = \tfrac{1}{2}(X_{(n)} + X_{(1)}) . \tag{9.62}$$

- Les **étendues** E_m de l'échantillon sont des statistiques définies par :

$$E_m = X_{(n-m+1)} - X_{(m)} . \tag{9.63}$$

  On réserve quelquefois le nom d'*empan* à la statistique E_1, c'est-à-dire à la portion de l'axe des x comprise entre la plus grande et la plus petite valeur de l'échantillon.

- Les **écarts** et les **écarts minimaux**. Un écart A_{ij} est une statistique égale à $X_{(j)} - X_{(i)}$. Un écart minimal d'ordre r est égal à :

$$A_r = \min_{|j-i|=r} A_{ij}, \quad r \leq n-1 . \tag{9.64}$$

  Parmi ceux-ci on distingue A_1, qui peut servir à approximer le mode de la population parente.

- La **médiane** est un point $Q_{0.5}$ sur l'axe des réalisations (axe des x) où il y a autant de valeurs X_i qui lui sont strictement inférieures que de valeurs qui lui sont strictement supérieures. Avec cette définition, la valeur de la

médiane $Q_{0.5}$ dépend de la parité de l'échantillon. Si l'échantillon est de taille impaire $n = 2p + 1$ on aura de façon unique $Q_{0.5} = X_{(p+1)}$. En revanche, si l'échantillon est de taille paire $n = 2p$, la médiane n'est pas définie de façon univoque ; en effet tout point entre $X_{(p)}$ et $X_{(p+1)}$ satisfait à la définition, et, par convention on prendra $Q_{0.5} = \frac{1}{2}(X_{(p+1)} + X_{(p)})$. Résumons cette définition par la formule suivante :

$$Q_{0.5} = \begin{cases} X_{(p+1)} & \text{si } n = 2p+1, \\ \frac{1}{2}(X_{(p+1)} + X_{(p)}) & \text{si } n = 2p. \end{cases} \tag{9.65}$$

- Les **quantiles** de l'échantillon. Un quantile $Q_{\alpha,n}$ d'un n-échantillon est défini par la variable aléatoire $X_{(n-\lfloor n\alpha \rfloor)}$, où $\lfloor n\alpha \rfloor$ désigne la partie entière de $n\alpha$. Le quantile $Q_{\alpha,n}$ est, avec cette définition, la variable ordonnée $X_{(i)}$ pour laquelle on a au plus $100\alpha\%$ de l'échantillon ayant une valeur strictement supérieure à $X_{(i)}$. Rappelons que le quantile x_α d'une population de loi $F(x)$, est défini par l'équation $F(x_\alpha) = 1 - \alpha$. Si F est continue, ce que nous supposerons, x_α est trouvé comme solution unique de $x_\alpha = F^{-1}(1 - \alpha)$.

▶**Exemple 9.7.** *Quantile d'un échantillon de petite taille.* Soit un échantillon ordonné de taille 5. L'indice du quantile d'ordre 0.25 est égal à $5 - \lfloor 0.25 \times 5 \rfloor = 4$, et donc $Q_{0.25,5} = X_{(4)}$. Considérons les 5 nombres issus d'une loi normale réduite :

$$x_{(1)} = -0.576\,, \quad x_{(2)} = -0.408\,, \quad x_{(3)} = 0.520\,, \quad x_{(4)} = 0.621\,, \quad x_{(5)} = 0.872\,.$$

Avec notre définition on a $Q_{0.25} = 0.621$.

Il existe bien d'autres statistiques basées sur les variables ordonnées. Toutes, comme celles que nous venons de voir, sont d'une extrême importance, en particulier dans les théories de l'estimation non paramétrique et de l'estimation fiable.

9.8 Exercices et problèmes.

Exercice 9.1. *Loi du minimum.* Trouver la fonction de répartition du minimum $X_{(1)}$ d'un échantillon de taille n, en effectuant d'abord le changement de variable $y = -x$ puis $x = -y$.

Exercice 9.2. *Fonction de répartition de la plus petite et de la plus grande variable aléatoire d'un couple.* Soit un couple de variables aléatoires (X, Y) de fonction de répartition $F(x, y)$ et de densité $f(x, y)$. Trouver l'expression de la fonction de répartition et de la densité de probabilité des variables aléatoires $U = \min(X, Y)$ et $V = \max(X, Y)$ dans le cas général où les variables X et Y sont dépendantes.

Donner ces expressions dans le cas où X et Y sont indépendantes et dans le cas où les lois marginales du couple sont identiques. Vérifier que ce dernier résultat est compatible avec les formules (9.12a) et (9.12b).

Exercice 9.3. *Variables uniformes.* Soit $(X_1, \ldots, X_n)$ un échantillon i.i.d de population parente uniforme sur $[0, 1]$. Montrer que les moyennes des variables ordonnées partagent l'intervalle $[0, 1]$ en $n + 1$ intervalles égaux, c'est-à-dire que :

$$\mathrm{E}\{X_{(k)}\} = \frac{k}{n+1}\,.$$

Soit $X_{(n)}$ la valeur maximum de l'échantillon, montrer que sa variance est égale à :

$$\mathrm{Var}(X_{(n)}) = \frac{n}{(n+1)^2(n+2)}\,.$$

Montrer que la variable aléatoire $n(X_{(n)}-1)$ converge en loi, lorsque $n \to \infty$, vers une variable aléatoire Y qui suit une loi de Weibull de fonction de répartition $G_{2,1}(y)$ telle que :

$$n(X_{(n)}-1) \xrightarrow{\text{loi}} Y\,, \quad \Pr\{Y \le y\} = G_{2,1}(y) = \begin{cases} e^y\,, & \text{si } y < 0\,; \\ 1\,, & \text{si } y \ge 0\,. \end{cases}$$

Note : Ce résultat est un cas particulier d'un théorème général qui établit l'existence de deux nombres a_n et b_n tels que la variable aléatoire $a_n(X_{(n)}-b_n)$ tend en loi vers une variable aléatoire dont la fonction de répartition appartient à seulement trois types (voir Fisher & Tippett (1928) [24] et Gnedenko (1943) [26] voir aussi Kendall & Stuart [40] chap 14.13–14.18).

Exercice 9.4. *La courbe de Quetelet.* Ramassez, un jour d'automne, une poignée de feuilles de saule tombées de l'arbre. Rejettez les feuilles « anormales » (celles qui sont abimées, tordues...) et triez celles qui restent par ordre de taille croissante en les disposant les unes à côté des autres. Si toutes les extrémités sont alignées d'un côté, que représente la courbe formée par les autres extrémités ? (van der Waerden § 15 [69].)

Exercice 9.5. *Fonction de vraisemblance d'un échantillon uniforme.* La population parente d'un échantillon i.i.d de taille n : $(X_1, \ldots, X_n)$ est la loi uniforme entre 0 et θ ($X_i = \mathcal{U}(0,\theta)$). Le paramètre θ est inconnu.

Donner l'expression de la fonction de vraisemblance de cet échantillon et tracer son graphe en fonction de θ pour une réalisation de l'échantillon que l'on simulera à l'aide d'un programme.

Même question mais pour des X_i suivant la loi de Cauchy.

Problème 9.6. *Détermination de la distance des galaxies en fonction de la taille apparente des régions HII.* On propose d'estimer la distance d'une galaxie en fonction du diamètre angulaire de la 3^{e} plus grande région HII visible. Déterminer la loi de la distance ainsi trouvée, étudier en particulier la sensibilité de la loi au nombre total de régions HII et aux paramètres de la loi donnant la distribution des diamètres réels. Dire enfin s'il ne vaudrait pas mieux déterminer la distance des galaxies à partir de la 5^{e}, de la 2^{e}, ou de tout autre rang du diamètre de la région HII. On consultera au préalable l'article de Hodge (1983) [32].

Chapitre 10

Echantillons issus d'une population normale.

Nous ne considérerons dans ce chapitre que des n-échantillons i.i.d extraits d'une population parente normale. Le cas normal bénéficie déjà d'une position privilégiée grâce au théorème central limite, mais en outre, on peut déterminer les lois exactes suivies par la moyenne empirique $\overline{X}_n$, la variance empirique S_n^2 et on peut montrer que ces deux variables aléatoires sont indépendantes. Ces propriétés, d'une grande importance pratique, constituent le théorème de Fisher.

Nous rappelons tout d'abord quelques résultats concernant les caractéristique numériques de la moyenne et de la variance empirique.

Espérance et variance de la moyenne empirique. La moyenne empirique, $\overline{X}_n$ d'un échantillon extrait d'une population quelconque de moyenne μ possède une espérance égale à cette moyenne :

$$\mathrm{E}\{\overline{X}_n\} = \mu\,. \tag{10.1}$$

De plus si l'échantillon est i.i.d la loi forte des grands nombres nous dit que $\overline{X}_n$ converge presque-sûrement vers μ et si la variance σ^2 de la population parente existe on a :

$$\mathrm{Var}(\overline{X}_n) = \frac{\sigma^2}{n}\,. \tag{10.2}$$

Espérance et variance de la variance empirique. Dans le cas où les X_i suivent une loi normale, on a $\mu_2 = \sigma^2$, $\mu_4 = 3\sigma^4$ et il vient :

$$\mathrm{E}\{S'^2\} = \sigma^2(1 - \frac{1}{n})\,, \tag{10.3}$$

$$\mathrm{Var}(S'^2) = \frac{2\sigma^4}{n}(1 - \frac{1}{n})\,. \tag{10.4}$$

Pour la variance modifiée $S^2 = nS'^2/(n-1)$ on a :

$$\mathrm{E}\{S^2\} = \sigma^2 \,, \tag{10.5}$$

$$\mathrm{Var}(S^2) = \frac{2\sigma^4}{n-1} \,. \tag{10.6}$$

Nous verrons plus loin qu'une des conséquences du théorème de Fisher est que :

$$\mathrm{Cov}(\overline{X}_n, S^2) = 0 \,. \tag{10.7}$$

En ce qui concerne la convergence en loi de la moyenne et de la variance empiriques, le théorème central limite nous fournit un résultat asymptotique mais celui de Fisher nous donne beaucoup mieux, il nous donne des expressions exactes pour tout n.

10.1 Le théorème de Fisher.

Théorème 10.1. *(Fisher.) Si un n-échantillon i.i.d $(X_1, \ldots, X_n)$ est issu d'une loi normale de moyenne μ et de variance σ^2, alors :*

1. *la moyenne empirique $\overline{X}_n$ suit une loi normale de moyenne μ et de variance σ^2/n.*

2. *la statistique $(n-1)S_n^2/\sigma^2$ suit une loi du χ^2 à $n-1$ degrés de liberté.*

3. *la moyenne empirique $\overline{X}_n$ et la variance empirique S_n^2 sont des variables aléatoires indépendantes.*

La démonstration de ce résultat remarquable va occuper les trois prochaines sections.

10.1.1 Loi suivie par la moyenne $\overline{X}_n$ d'un échantillon normal.

Nous noterons donc $\overline{X}_n$ la moyenne du n-échantillon i.i.d : $(X_1, \ldots, X_n)$ de population parente normale $X_i = \mathcal{N}(\mu, \sigma^2)$. Son expression $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ nous montre que cette variable aléatoire est la somme de n variables aléatoires normales, indépendantes. La loi suivie par $\overline{X}_n$ est donc également une loi normale, de moyenne $\mathrm{E}\{\overline{X}_n\} = \mu$ et de variance $\mathrm{Var}(\overline{X}_n) = \sigma^2/n$. On a : $\overline{X}_n = \mathcal{N}(\mu, \sigma^2/n)$, ou encore en réduisant la variable $\overline{X}_n$:

$$\sqrt{n}\frac{\overline{X}_n - \mu}{\sigma} = \mathcal{N}(0,1) \,. \tag{10.8}$$

Nous savions d'après le théorème central limite que le membre de gauche de l'équation précédente devait *tendre* en loi vers $\mathcal{N}(0,1)$ mais ici, dans le cas particulier d'une population parente normale, il est *égal* à la loi $\mathcal{N}(0,1)$.

10.1.2 Loi suivie par la variance modifiée S_n^2 d'un échantillon normal.

Nous prendrons S_n^2 comme définition de la variance de l'échantillon, son expression est :

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2\,, \tag{10.9}$$

$$S_n^2 = \frac{1}{n-1}\Big[\sum_{i=1}^{n} X_i^2 - \frac{1}{n}\Big(\sum_{i=1}^{n} X_i\Big)^2\Big]\,, \tag{10.10}$$

cette écriture met en évidence que S_n^2 est la somme de variables aléatoires *dépendantes*. Introduisons les variables réduites $X_i' = (X_i - \mu)/\sigma$, elles sont indépendantes et suivent une loi normale réduite : $X_i' = \mathcal{N}(0,1)$. Portons ces nouvelles variables dans l'équation (10.10). Il vient :

$$S_n^2 = \frac{\sigma^2}{n-1}\Big[\sum_{i=1}^{n} {X_i'}^2 - \frac{1}{n}\Big(\sum_{i=1}^{n} X_i'\Big)^2\Big]\,. \tag{10.11}$$

Cherchons à déterminer la fonction de répartition de l'expression entre crochets, et en anticipant un peu la notation sur le résultat, posons :

$$\chi^2 = \sum_{i=1}^{n} {X_i'}^2 - \frac{1}{n}\Big(\sum_{i=1}^{n} X_i'\Big)^2 = (n-1)\frac{S_n^2}{\sigma^2}\,. \tag{10.12}$$

Les variables aléatoires X_i' étant indépendantes, leur densité de probabilité conjointe f_n est égale au produit de leurs densités. Il vient :

$$\begin{aligned} f_n(x_1', \ldots, x_n') &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\big(-\tfrac{1}{2}{x_i'}^2\big)\,, \\ &= \frac{1}{(2\pi)^{n/2}} \exp\Big(-\tfrac{1}{2}\sum_{i=1}^{n} {x_i'}^2\Big)\,. \end{aligned}$$

La fonction de répartition F de la variable aléatoire χ^2, $F(u) = \Pr\{\chi^2 \le u\}$, $u \ge 0$, s'obtient comme fonction de répartition conditionnelle des X_i' :

$$F(u) = \int\!\cdots\!\int_{\chi^2 \le u} f_n(x_1', \ldots, x_n')\, dx_1' \cdots dx_n'\,. \tag{10.13}$$

La constante de normalisation s'obtient en intégrant f_n sur tout $\mathbb{R}^n$, elle est donc égale à 1. Le bord du domaine d'intégration a pour équation $\chi^2 = u$. C'est une forme quadratique homogène d'équation $u = \sum {x_i'}^2 - \frac{1}{n}(\sum x_i')^2$. Cette forme quadratique est, nous allons le voir, l'équation d'un hyper-cylindre dans $\mathbb{R}^n$. La matrice caractéristique $\boldsymbol{G}$ de la forme quadratique u a pour éléments $g_{ij} = \frac{1}{2}\partial_{ij}^2 u = \delta_{ij} - u_i u_j$, où δ_{ij} est le symbole de Kronecker, et les u_i les composantes du vecteur $\boldsymbol{u} = (n^{-\frac{1}{2}}, \ldots, n^{-\frac{1}{2}})$. Dans $\mathbb{R}^3$ la matrice $\boldsymbol{G}$ serait égale à l'expression suivante :

$$\boldsymbol{G} = \begin{pmatrix} \frac{2}{3} & -\frac{1}{3} & -\frac{1}{3} \\ -\frac{1}{3} & \frac{2}{3} & -\frac{1}{3} \\ -\frac{1}{3} & -\frac{1}{3} & \frac{2}{3} \end{pmatrix}\,. \tag{10.14}$$

Ecrivons $\boldsymbol{G}$ sous forme matricielle. On a $\boldsymbol{G} = \boldsymbol{I} - \boldsymbol{u}\boldsymbol{u}^t$, avec $\boldsymbol{u}^t\boldsymbol{u} = 1$. La matrice $\boldsymbol{G}$ est un projecteur, qui projette $\mathbb{R}^n$ sur un sous-espace de dimension $n-1$ le long de la direction indiquée par le vecteur $\boldsymbol{u}$. En tant que projecteur on a $\boldsymbol{G}^2 = \boldsymbol{G}$ et ses valeurs propres sont 0 ou 1. La valeur propre $\lambda_1 = 0$ correspond au vecteur propre $\boldsymbol{u}$, et la valeur propre 1 correspond à un sous-espace propre, dont la dimension est donnée par la trace de la matrice qui est ici égale à $n-1$. Effectuons un changement de variables unitaire (une rotation) où les nouvelles coordonnées y_i sont telles que l'axe des y_1 correspond au vecteur $\boldsymbol{u}$. Dans cette nouvelle base, la matrice $\boldsymbol{G}$ prend sa forme diagonale $\boldsymbol{G} = \mathrm{diag}(0, 1, \ldots, 1)$ il s'ensuit que la forme quadratique (10.12) et l'équation du bord du domaine d'intégration s'écrivent :

$$\chi^2 = \sum_{i=1}^{n} \lambda_i Y_i^2 = \sum_{i=2}^{n} Y_i^2 \,, \quad u = \sum_{i=2}^{n} y_i^2 \,. \tag{10.15}$$

La dernière équation est bien celle d'un cylindre de section circulaire et ayant pour axe : l'axe des y_1. La figure 10.1 donne une interprétation géométrique du calcul du χ^2.

Evaluons maintenant l'intégrale (10.13), en remplaçant f_n par sa valeur exprimée avec les nouvelles variables, comme il s'agit d'un changement de variables unitaire, le jacobien est égal à un. Il vient, en extrayant de f_n la dépendance par rapport à la première variable :

$$F(u) = (2\pi)^{-\frac{n-1}{2}} \int_{\chi^2 \le u} \cdots \int \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2} y_1^2) \, \exp\Big(-\frac{1}{2} \sum_{i=2}^{n} y_i^2\Big) \, dy_1 \cdots dy_n \,.$$

Comme d'après l'équation (10.15) u ne dépend pas de y_1, on peut intégrer y_1 de $-\infty$ à ∞, et il ne plus reste plus qu'à évaluer :

$$F(u) = (2\pi)^{-\frac{n-1}{2}} \int_{\chi^2 \le u} \cdots \int \exp\Big(-\frac{1}{2} \sum_{i=2}^{n} y_i^2\Big) \, dy_2 \cdots dy_n \,.$$

Cette dernière équation est la fonction de répartition de la somme des carrés de $n-1$ variables aléatoires normales réduites indépendantes $\chi^2 = \sum_{i=2}^{n} Y_i^2$, ce qui suffit à montrer, d'après les résultats du chapitre 6.3.9 (page 101), que χ^2 suit une loi du χ^2 à $n-1$ degrés de liberté. On a alors :

$$F(u) = \frac{1}{2\Gamma\left(\frac{n-1}{2}\right)} \int_0^u e^{-\frac{t}{2}} \left(\frac{t}{2}\right)^{\frac{n-1}{2}-1} dt \,. \tag{10.16}$$

La variable aléatoire $\chi^2 = (n-1)S_n^2/\sigma^2$ admet donc la densité de probabilité :

$$f(u) = \frac{1}{2\Gamma\left(\frac{n-1}{2}\right)} e^{-\frac{u}{2}} \left(\frac{u}{2}\right)^{\frac{n-1}{2}-1} \qquad u \ge 0 \,, \tag{10.17}$$

La loi du χ^2_{n-1} a pour moyenne $n-1$ et pour variance $2(n-1)$. Il vient donc :

$$\mathrm{E}\left\{(n-1)\frac{S_n^2}{\sigma^2}\right\} = n-1, \qquad \mathrm{Var}\Big((n-1)\frac{S_n^2}{\sigma^2}\Big) = 2(n-1), \tag{10.18}$$

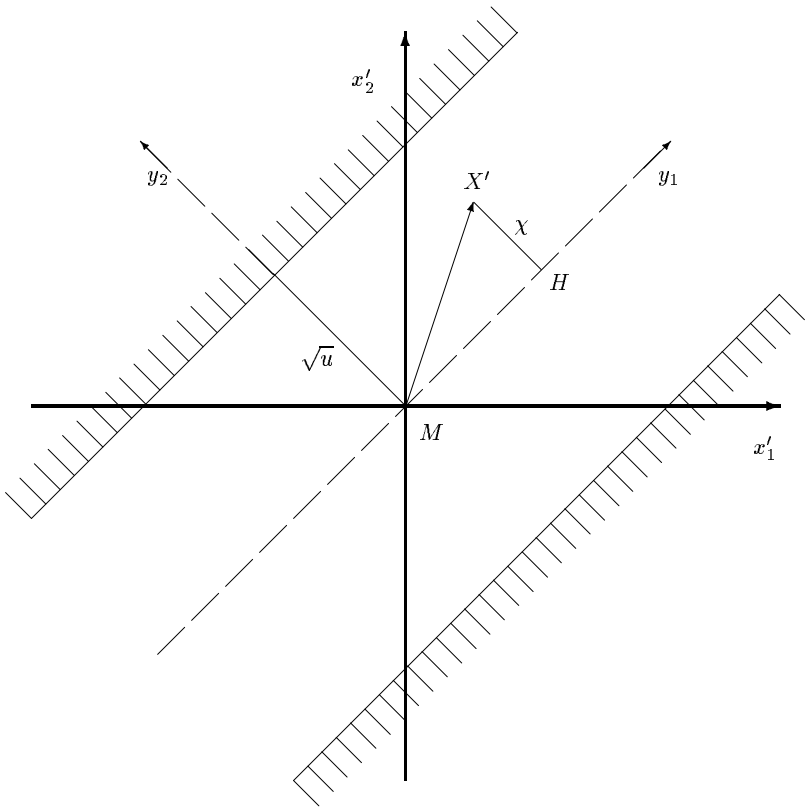


FIG. 10.1: *Domaine d'intégration pour le calcul de la fonction de répartition de la variance empirique S^2, par l'intermédiaire de $\chi^2 = (n-1)S^2/\sigma^2$. Le domaine d'intégration est un cylindre de rayon $\sqrt{u}$, qui est ici, dans $\mathbb{R}^2$, réduit à deux droites parallèles à la 1re bissectrice des axes. On a $\Pr\{\chi^2 \leq u\} = F(u)$, qui est une loi du χ^2 à $n-1$ degrés de liberté (ici $n = 2$). Le point X' représente une issue quelconque du couple de variables aléatoires indépendantes réduites et centrées: (X'_1, X'_2). On montre que la variable χ^2 est égale au carré de la distance du point X' à l'axe du cylindre. Le segment de droite MH, est égal à la variable $\sqrt{n}(\bar{X}_n - \mu)/\sigma$ qui suit une loi normale réduite. Le produit par $\sqrt{n-1}$ de la cotangente de l'angle $\widehat{X'MH}$ suit la loi de Student. Ce résultat se généralise à $\mathbb{R}^n$.*

ce qui nous permet de retrouver l'espérance et la variance de la variance de l'échantillon modifiée S_n^2 :

$$\mathrm{E}\left\{S_n^2\right\} = \sigma^2, \quad \mathrm{Var}(S_n^2) = \frac{2\sigma^4}{n-1}\cdot \tag{10.19}$$

A partir de ces expressions on trouve pour la variance $S_n'^2 = (n-1)S_n^2/n$:

$$\mathrm{E}\{S_n'^2\} = \frac{n-1}{n}\sigma^2, \quad \mathrm{Var}(S_n'^2) = 2\frac{n-1}{n^2}\sigma^4. \tag{10.20}$$

10.1.3 Indépendance de $\overline{X}_n$ et S_n^2.

Calculons la fonction de répartition conjointe des deux variables aléatoires $\overline{Y}_n = \sqrt{n}(\overline{X}_n - \mu)/\sigma$ et $\chi^2 = (n-1)S_n^2/\sigma^2$. Elle s'obtient comme fonction de répartition conditionnelle dans $\mathbb{R}^n$, calculée à partir de la densité f_n de l'expression (10.13) :

$$F(u,v) = F(x'_1, \ldots, x'_n | \chi^2 \leq u, \overline{Y_n} \leq v)\,, \tag{10.21}$$

$$F(u,v) = \int_{\chi^2 \leq u, \overline{Y}_n \leq v} \cdots \int f_n(x'_1, \ldots, x'_n)\, dx'_1 \cdots dx'_n\,, \tag{10.22}$$

avec le même changement de variable unitaire qu'au paragraphe précédent, où, de même que les produits scalaires, les formes quadratiques χ^2 et u restent invariantes. La borne d'intégration :

$$v = \overline{Y}_n = \sqrt{n}\frac{1}{n}\sum_{i=1}^{n} x'_i, \tag{10.23}$$

est le produit scalaire des vecteurs $v = (1/\sqrt{n}, \ldots, 1/\sqrt{n}) \cdot (x'_1, \ldots, x'_n)$, qui dans la transformation deviennent $v = (1, 0, \ldots, 0) \cdot (y_1, y_2, \ldots, y_n) = y_1$, d'où $v = y_1$. Remplaçons ce résultat dans (10.22), il vient :

$$F(u,v) = (2\pi)^{-\frac{n}{2}} \int_{-\infty}^{v} e^{-\frac{1}{2}y_1^2}\, dy_1 \int_{\chi^2 \leq u} \cdots \int e^{-\frac{1}{2}\chi^2}\, dy_2 \cdots dy_n\,. \tag{10.24}$$

Cette intégrale se présente bien sous la forme du produit des deux fonctions de répartition des variables $\overline{Y}_n$ et χ^2 :

$$F(u,v) = F_{\overline{Y}_n}(v) F_{\chi^2}(u)\,, \tag{10.25}$$

ce qui démontre que les variables $\sqrt{n}(\overline{X}_n - \mu)/\sigma$ et $(n-1)S_n^2/\sigma^2$ sont indépendantes, et que par conséquent $\overline{X}_n$ et S_n^2 le sont également. Cela termine la démonstration du théorème de Fisher. Notons que sur la figure 10.1, la variable $Y_1 = \sqrt{n}(\overline{X_n} - \mu)/\sigma$ est égale à la longueur du segment de droite MH comptée sur l'axe du cylindre.

10.2 La loi de « Student ».

Nous savons maintenant que $(n-1)S_n^2/\sigma^2$ et $\overline{X}_n$ sont des variables aléatoires indépendantes suivant respectivement une loi du χ^2_{n-1} à $n-1$ degrés de liberté et une loi normale de moyenne μ et de variance σ^2/n. La loi de S_n^2 dépend du paramètre σ^2 que S_n^2 prétend estimer, ce qui n'est pas surprenant, mais la loi de $\overline{X}_n$ dépend de deux paramètres : μ, que $\overline{X}_n$ veut estimer mais aussi de σ^2, ce qui est une difficulté. Dans la pratique, on peut tenter d'approximer la variable réduite $\sqrt{n}(\overline{X}_n - \mu)/\sigma$ qui suit la loi normale réduite, par la variable aléatoire T_n où l'on remplace σ par $S_n = \sqrt{S_n^2}$, soit donc :

$$T_n = \sqrt{n}\frac{\overline{X}_n - \mu}{S_n}\,. \tag{10.26}$$

Pour résoudre le problème de l'estimation de μ il faut calculer la loi suivie par T_n, et c'est ce calcul qui constitue le célèbre travail de « Student [1] ». Tentons d'exprimer T_n en fonction de variables obéissant à des lois connues :

$$\begin{aligned} T_n &= \sqrt{n}\frac{\overline{X}_n - \mu}{S_n}\,, \\ &= \sqrt{n}\frac{\overline{X}_n - \mu}{\sigma}\frac{\sigma}{S_n}\,, \\ &= \sqrt{n-1}\Big(\sqrt{n}\frac{\overline{X}_n - \mu}{\sigma}\Big)\Big(\sqrt{\frac{\sigma^2}{(n-1)S_n^2}}\Big)\,. \end{aligned}$$

Avec les notations précédentes on a :

$$T_n = \sqrt{n-1}\,\frac{\overline{Y}_n}{\sqrt{\chi^2}}\,. \tag{10.27}$$

La loi de T_n apparaît alors comme le produit par $\sqrt{n-1}$ du quotient d'une variable aléatoire suivant la loi normale réduite par la racine carrée d'une variable aléatoire suivant la loi du χ^2_{n-1}. On cherche la loi de T_n, c'est-à-dire $F_{T_n}(t) = \Pr\{T_n \leq t\}$; les variables $\overline{Y}_n$ et χ^2 étant indépendantes, cette probabilité est calculée à partir du produit des densités de $\overline{Y}_n$ et de χ^2. Il vient :

$$F_{T_n}(t) = \iint\limits_{T_n \leq t} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2}\,\frac{1}{2\Gamma\left(\frac{n-1}{2}\right)}\left(\frac{u}{2}\right)^{\frac{n-1}{2}-1} e^{-\frac{1}{2}u}\,du dy\,. \tag{10.28}$$

C'est cette intégrale qu'il faut calculer afin de résoudre notre problème. Posons $f = n-1$, l'intégrale s'écrit alors :

$$F_{T_n}(t) = \frac{2^{-\frac{f+1}{2}}}{\sqrt{\pi}\,\Gamma\left(\frac{f}{2}\right)} \iint\limits_{y\sqrt{f/u} \leq t} e^{-\frac{1}{2}y^2} u^{\frac{f}{2}-1} e^{-\frac{1}{2}u}\,du dy\,. \tag{10.29}$$

Posons α égal à la constante située hors de l'intégrale, et en remarquant que le domaine d'intégration est borné par $u > 0$ et $y = t\sqrt{u/f}$, on obtient :

$$F_{T_n}(t) = \alpha \int_0^{\infty} du \int_{-\infty}^{t\sqrt{\frac{u}{f}}} e^{-\frac{1}{2}y^2} u^{\frac{f}{2}-1} e^{-\frac{1}{2}u}\,dy\,. \tag{10.30}$$

Effectuons le changement de variable $x = y\sqrt{f/u}$. Il vient :

$$\begin{aligned} F_{T_n}(t) &= \alpha \int_0^{\infty} du \int_{-\infty}^{t} e^{-\frac{x^2 u}{2f}} u^{\frac{f}{2}-1} e^{-\frac{1}{2}u}\,dx\,, \\ &= \frac{\alpha}{\sqrt{f}} \int_0^{\infty} du \int_{-\infty}^{t} u^{\frac{f}{2}-1} e^{-\frac{1}{2}u\left(1+\frac{x^2}{f}\right)}\,dx\,, \\ &= \frac{\alpha}{\sqrt{f}} \int_{-\infty}^{t} dx \int_0^{\infty} u^{\frac{f}{2}-1} e^{-\frac{1}{2}u\left(1+\frac{x^2}{f}\right)}\,du\,. \end{aligned}$$

1. Voir "On the probable error of mean", Student (1908) [68].

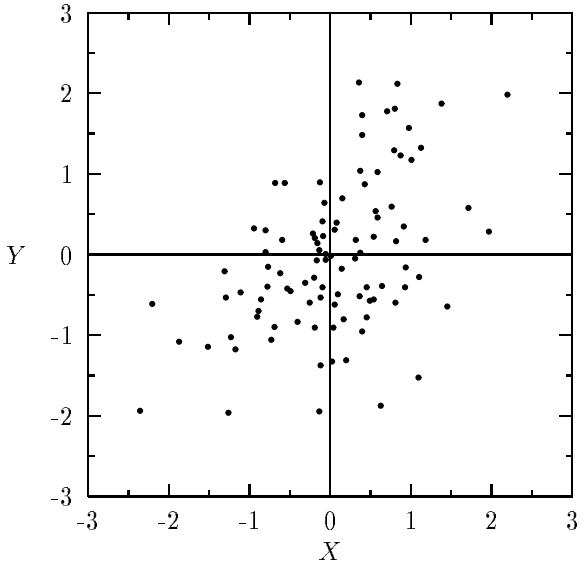


FIG. 10.2: *Représentation graphique d'un échantillon normal 2D de taille 100. La population parente normale est de moyenne nulle, $\mu_1 = \mu_2 = 0$, de variances unité $\sigma_1^2 = \sigma_2^2 = 1$ et de coefficient de corrélation $\rho = 0.5$.*

On a pu inverser l'ordre des intégrations car l'intégrale double (10.28) existe. Sachant que $\int_0^\infty u^{\nu-1} e^{-\mu u}\, du = \mu^{-\nu}\Gamma(\nu)$ pour $\mu > 0$ et $\nu > 0$, il vient :

$$F_{T_n}(t) = \frac{\alpha}{\sqrt{f}} \int_{-\infty}^{t} dx \left[\frac{1}{2}\Big(1 + \frac{x^2}{f}\Big)\right]^{-\frac{f+1}{2}} \Gamma\Big(\frac{f+1}{2}\Big)\,.$$

On trouve finalement la fonction de répartition de T_n, en remplaçant la constante α par sa valeur :

$$F_{T_n}(t) = \frac{\Gamma\left(\frac{f+1}{2}\right)}{\sqrt{\pi f}\,\Gamma\left(\frac{f}{2}\right)} \int_{-\infty}^{t} dx \left(1 + \frac{x^2}{f}\right)^{-\frac{f+1}{2}}, \tag{10.31}$$

et sa densité de probabilité :

$$f_{T_n}(t) = \frac{\Gamma\left(\frac{f+1}{2}\right)}{\sqrt{\pi f}\,\Gamma\left(\frac{f}{2}\right)} \left(1 + \frac{t^2}{f}\right)^{-\frac{f+1}{2}}, \tag{10.32}$$

ce qui montre que la variable aléatoire T_n suit une loi de Student à $f = n - 1$ degrés de liberté.

10.3 Echantillons issus d'une loi normale 2D.

Nous étudions maintenant un n-échantillon $((X_1, Y_1), \ldots, (X_n, Y_n))$ formé de couples de variables aléatoires (X, Y) suivant la loi normale 2D. Il est pratique de représenter, comme sur la figure 10.2, une réalisation de cet échantillon par des points répartis sur un plan. Nous supposons toujours que les couples (X_i, Y_i) sont indépendants (échantillon i.i.d) mais, en revanche, les variables X_i et Y_i

ne sont pas nécessairement indépendantes. La densité de la loi normale 2D est donnée par l'expression :

$$f(x,y) = \frac{1}{2\pi\sigma_1\sigma_2(1-\rho^2)^{\frac{1}{2}}} \exp\Big\{-\frac{1}{2(1-\rho^2)} \Big[\frac{(x-\mu_1)^2}{\sigma_1^2} - 2\rho\frac{(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2} + \frac{(y-\mu_2)^2}{\sigma_2^2}\Big]\Big\}, \quad (10.33)$$

de sorte que la densité du n-échantillon pour la réalisation $((x_1,y_1),\ldots,(x_n,y_n))$, est donnée par l'expression :

$$f(x_1,y_1,\ldots,x_n,y_n) = \frac{1}{(2\pi\sigma_1\sigma_2)^n(1-\rho^2)^{\frac{n}{2}}} \exp\Big\{-\frac{1}{2(1-\rho^2)} \Big[\frac{1}{\sigma_1^2}\sum_{i=1}^n (x_i-\mu_1)^2 - \frac{2\rho}{\sigma_1\sigma_2}\sum_{i=1}^n (x_i-\mu_1)(y_i-\mu_2) + \frac{1}{\sigma_2^2}\sum_{i=1}^n (y_i-\mu_2)^2\Big]\Big\}. \quad (10.34)$$

On définit, à l'aide des moments échantillonnés de la loi 2D, les cinq statistiques suivantes :

$$\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i\,, \qquad \overline{Y} = \frac{1}{n}\sum_{i=1}^n Y_i\,, \quad (10.35)$$

$$S_1'^2 = \frac{1}{n}\sum_{i=1}^n (X_i-\overline{X})^2\,, \qquad S_2'^2 = \frac{1}{n}\sum_{i=1}^n (Y_i-\overline{Y})^2\,, \quad (10.36)$$

$$R = \frac{\frac{1}{n}\sum_{i=1}^n (X_i-\overline{X})(Y_i-\overline{Y})}{S_1'S_2'}\,. \quad (10.37)$$

La fonction de répartition F_5 de ces cinq variables aléatoires $\overline{X},\overline{Y},S_1',S_2',R$ est trouvée par intégration de f_n sur le domaine $\mathcal{D}$: $\{\overline{X}\le\overline{x},\overline{Y}\le\overline{y},S_1'\le s_1',S_2'\le s_2',R\le r\}$:

$$F_5(\overline{x},\overline{y},s_1',s_2',r) = \int\cdots\int_{\mathcal{D}} f_n(x_1,y_1,\ldots,x_n,y_n)\,dx_1dy_1\cdots dx_ndy_n\,. \quad (10.38)$$

L'intégration se conduit de façon analogue à celle exposée dans le théorème de Fisher, par diagonalisation des formes quadratiques entrant dans f_n. Le même Fisher (1915) [22] a donné l'expression de la densité f_5 et il a montré qu'elle pouvait être séparée en deux expressions indépendantes : $dF_5 = dF_m dF_v$, avec :

$$dF_m(\overline{x},\overline{y}) = \frac{1}{2\pi\sigma_1\sigma_2(1-\rho^2)^{\frac{1}{2}}} \exp\Big\{-\frac{n}{2(1-\rho^2)} \Big[\frac{(\overline{x}-\mu_1)^2}{\sigma_1^2} - 2\rho\frac{(\overline{x}-\mu_1)(\overline{y}-\mu_2)}{\sigma_1\sigma_2} + \frac{(\overline{y}-\mu_2)^2}{\sigma_2^2}\Big]\Big\}\,d\overline{x}d\overline{y}\,, \quad (10.39a)$$

$$dF_v(s_1',s_2',r) = \frac{n^{n-1}{s_1'}^{n-2}{s_2'}^{n-2}(1-r^2)^{\frac{1}{2}(n-4)}}{\pi\sigma_1^{n-1}\sigma_2^{n-1}(1-\rho^2)^{\frac{1}{2}(n-2)}\Gamma(n-2)} \exp\Big\{-\frac{n}{2(1-\rho^2)}\Big[\frac{{s_1'}^2}{\sigma_1^2} - 2\rho\frac{rs_1's_2'}{\sigma_1\sigma_2} + \frac{{s_2'}^2}{\sigma_2^2}\Big]\Big\}\,ds_1'ds_2'dr\,. \quad (10.39b)$$

Cette factorisation montre que le couple $(\overline{X}, \overline{Y})$ d'une part, et le triplet (S'_1, S'_2, R) d'autre part, sont indépendants. De plus, l'équation (10.39a) nous dit que le couple $\overline{X}, \overline{Y}$ suit une loi normale 2D de moyenne μ_1, μ_2 et de matrice des variances-covariances :

$$\boldsymbol{V} = \begin{pmatrix} \dfrac{\sigma_1^2}{n} & \rho\dfrac{\sigma_1\sigma_2}{n} \\ \rho\dfrac{\sigma_1\sigma_2}{n} & \dfrac{\sigma_2^2}{n} \end{pmatrix} . \tag{10.40}$$

Le couple $\overline{X}, \overline{Y}$ suit donc la même loi que la population parente : il possède la même moyenne, le même coefficient de corrélation, mais ses variances $\sigma_1^2/n, \sigma_2^2/n$ sont différentes ; elles tendent d'ailleurs vers 0 quand n tend vers l'infini. Les lois marginales de la loi normale 2D étant normales, $\overline{X}$ suit donc une loi normale de moyenne μ_1 et de variance σ_1^2/n et $\overline{Y}$ une loi normale de moyenne μ_2 et de variance σ_2^2/n.

Pour obtenir la loi du coefficient de corrélation empirique R, il suffit d'intégrer dF_v par rapport à s'_1 et s'_2 de zéro à l'infini. Nous reviendrons sur cette loi au chapitre 17.1.2 page 300.

10.4 Exercices et problèmes.

Exercice 10.1. *Théorème de Fisher.* Le but de cet exercice est de démontrer d'une autre façon la 3e partie du théorème de Fisher établissant l'indépendance de $\overline{X}_n$ et S_n^2.

Soit un échantillon i.i.d issu de la loi normale $\mathcal{N}(\mu, \sigma^2)$, on sait que la moyenne empirique $\overline{X}_n$ suit une loi $\mathcal{N}(\mu, \frac{\sigma^2}{n})$.

1. Démontrer que la variable aléatoire $X_i - \overline{X}_n$ suit une loi normale de moyenne nulle et de variance $(n-1)\sigma^2/n$, c'est-à-dire que :
$$X_i - \overline{X}_n = \mathcal{N}(0, \tfrac{n-1}{n}\sigma^2) .$$
2. Montrer que $\mathrm{Cov}(\overline{X}_n, X_i - \overline{X}_n) = 0$ et en déduire que les variables aléatoires $\overline{X}_n$ et $X_i - \overline{X}_n$ sont indépendantes.
3. Finalement en déduire que $\overline{X}_n$ et S_n^2 sont indépendantes.

Chapitre 11

L'estimation ponctuelle.

Les statistiques $g(X_1, \ldots, X_n)$ peuvent servir à approximer les paramètres inconnus des populations parentes. On appellera « *estimateur* » une statistique dont l'objectif consiste effectivement à estimer un de ces paramètres. Soit θ un tel paramètre : on notera $\hat{\theta}_n$ un estimateur de θ construit à l'aide d'un échantillon de taille n, et on appellera « *estimation* » une réalisation d'un estimateur. Il est traditionnel de noter avec le même symbole l'estimateur et l'estimation qui en résulte. Ainsi, quand on dira l'estimateur $\hat{\theta}_n$, il faudra entendre la statistique $\hat{\theta}_n(X_1, \ldots, X_n)$, et quand on dira l'estimation $\hat{\theta}_n$, il faudra entendre la réalisation $\hat{\theta}_n(x_1, \ldots, x_n)$. Un estimateur est une variable aléatoire, alors qu'une estimation est un nombre. On peut dire aussi, en termes plus imagés, qu'un estimateur est une « façon de faire », alors qu'une estimation en est le résultat.

Par exemple, la moyenne arithmétique de l'échantillon $(X_1, \ldots, X_n)$ est un estimateur de la moyenne μ de la population parente :

$$\hat{\mu}_n = M = \frac{1}{n} \sum_{i=1}^{n} X_i \,, \tag{11.1}$$

qui conduit à l'estimation :

$$\hat{\mu}_n = m = \frac{1}{n} \sum_{i=1}^{n} x_i \,. \tag{11.2}$$

Il est pratique d'élargir le cadre de l'estimation de θ par $\hat{\theta}$ à celui de l'estimation d'une fonction $t = \tau(\theta)$ par une statistique notée T, ou T_n lorsque la référence à la taille de l'échantillon s'avère nécessaire.

Pour être utiles, ces estimateurs doivent satisfaire un certain nombre de propriétés que nous allons définir ci-dessous.

11.1 La convergence.

On dit qu'un estimateur est convergent lorsque, la taille de l'échantillon tendant vers l'infini, il converge en probabilité vers la valeur qu'il prétend estimer, c'est-à-dire quand :

$$\forall \epsilon > 0, \quad \lim_{n \to \infty} \Pr\left\{ |\hat{\theta}_n - \theta| \geq \epsilon \right\} = 0 \,. \tag{11.3}$$

Plus généralement T est un estimateur convergent de $\tau(\theta)$ si :

$$\forall \epsilon > 0, \quad \lim_{n\to\infty} \Pr\{|T_n - \tau(\theta)| \geq \epsilon\} = 0 \,. \tag{11.4}$$

La propriété de convergence est absolument requise. Il n'est pas d'estimateur utile qui ne soit convergent.

11.1.1 Convergence de la moyenne et de la variance empirique.

Si une population possède une moyenne μ et une variance σ^2, alors la moyenne empirique M et la variance empirique S^2 d'un échantillon i.i.d issu de cette population sont des estimateurs convergents de μ et σ^2. Ce résultat est un des aspects de la loi des grands nombres étudiée au chapitre 7, donnons-en ici une démonstration pour la moyenne M.

D'après l'inégalité de Bienaymé-Tchébychev on a :

$$\Pr\left\{\left|\frac{M - \mathrm{E}\{M\}}{\mathrm{Var}(M)^{1/2}}\right| \geq k\right\} \leq \frac{1}{k^2} \tag{11.5}$$

Cette inégalité signifie que la variable aléatoire M ne s'écarte de sa moyenne de plus que k fois son écart type qu'avec une probabilité inférieure à $1/k^2$. Par ailleurs, nous avons déjà démontré que $\mathrm{E}\{M\} = \mu$ et que $\mathrm{Var}(M) = \sigma^2/n$, voir page 77, d'où il résulte que ces quantités existent si μ et σ^2 existent, il s'ensuit

$$\begin{aligned}
\Pr\left\{\left|\frac{M-\mu}{\sigma/\sqrt{n}}\right| \geq k\right\} &\leq \frac{1}{k^2} \\
\Pr\left\{|M-\mu| \geq k\frac{\sigma}{\sqrt{n}}\right\} &\leq \frac{1}{k^2}; \quad \text{on pose} \quad \epsilon = k\frac{\sigma}{\sqrt{n}} \\
\Pr\{|M-\mu| \geq \epsilon\} &\leq \frac{\sigma^2}{n\epsilon^2}
\end{aligned}$$

Cette dernière quantité tend vers 0 quand $n \to \infty$, ce qui montre que M est bien un estimateur convergent de μ.

11.2 L'absence de biais.

On définit le biais b d'un estimateur $\hat{\theta}_n$ de θ comme étant égal à la différence entre l'espérance mathématique de l'estimateur et le paramètre qu'il veut estimer, soit :

$$b(\hat{\theta}_n, \theta) = \mathrm{E}\{\hat{\theta}_n\} - \theta \tag{11.6}$$

Quand la confusion n'est pas possible on notera le biais simplement $b_n(\theta)$. On dit qu'un estimateur est « *non-biaisé* » ou « *absolument correct* » si $b_n(\theta) = 0$, c'est-à-dire si :

$$\mathrm{E}\{\hat{\theta}_n\} = \theta \quad \text{ou} \quad \mathrm{E}\{T_n\} = \tau(\theta) \,. \tag{11.7}$$

▶**Exemple 11.1.** La moyenne empirique M et la variance modifiée S^2 sont des estimateurs non-biaisés de la moyenne μ et de la variance σ^2 de la population parente (a condition toutefois que ces paramètres existent).

Si le biais tend vers 0 quand n tend vers l'infini, l'estimateur $\hat{\theta}_n$ est alors dit « *asymptotiquement correct* ». Un estimateur convergent est toujours asymptotiquement correct.

11.2.1 Biais de la variance d'un échantillon i.i.d.

Si la moyenne de la population est connue, la variance de l'échantillon est absolument correcte ; mais si on estime la moyenne de la population par la moyenne de l'échantillon, la variance de l'échantillon est seulement asymptotiquement correcte. Nous avons déjà mentionné ce fait (voir équation (9.52), page 186), démontrons-le ici de façon directe. La variance de l'échantillon est définie par :

$$S'^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - M)^2 \,, \tag{11.8}$$

avec comme d'habitude $M = \frac{1}{n}\sum_{i=1}^{n} X_i$. D'après la définition de la fonction de répartition empirique F_n, il vient :

$$M = \int x\, dF_n \quad \text{et} \quad S'^2 = \int (x - M)^2\, dF_n \,.$$

Appliquons le théorème de Huygens à la deuxième intégrale en considérant l'écart quadratique moyen autour de μ, il vient :

$$S'^2 = \int (x - \mu)^2\, dF_n - (\mu - M)^2 \,.$$

L'intégrale restante est la variance de l'échantillon par rapport à la moyenne de la population parente, pour un échantillon i.i.d cet estimateur est non-biaisé. Calculons à présent l'espérance mathématique de l'expression ci-dessus, en notant que pour un échantillon i.i.d $\mathrm{E}\{M\} = \mu$, il vient :

$$\begin{aligned}
\mathrm{E}\{S'^2\} &= \mathrm{E}\{\int (x - \mu)^2\, dF_n\} - \mathrm{E}\{(\mu - M)^2\} \,, \\
&= \sigma^2 - \mathrm{E}\{(M - \mathrm{E}\{M\})^2\} = \sigma^2 - \mathrm{Var}(M) \,, \\
&= \sigma^2 - \frac{\sigma^2}{n} \,.
\end{aligned}$$

On a ainsi obtenu l'espérance de la variance S'^2 de l'échantillon :

$$\mathrm{E}\{S'^2\} = \sigma^2 - \frac{\sigma^2}{n} \,. \tag{11.9}$$

Cette dernière équation montre que S'^2 est un estimateur biaisé de σ^2 ; son biais vaut $b_n(\sigma^2) = \mathrm{E}\{S'^2\} - \sigma^2 = -\sigma^2/n$. La variance de l'échantillon est donc systématiquement plus petite que la variance de la population.

Ce résultat était prévisible puisque M est le centre de gravité de l'échantillon, et que la variance S'^2 de l'échantillon en est le moment d'inertie par rapport à M. Or on sait, toujours d'après le théorème de Huygens,que le moment d'inertie est minimum lorsqu'il est calculé par rapport au centre de gravité. La vraie variance σ^2, qui est le monment d'inertie de l'échantillon par rapport à μ, ne peut être que plus petite que S'^2.

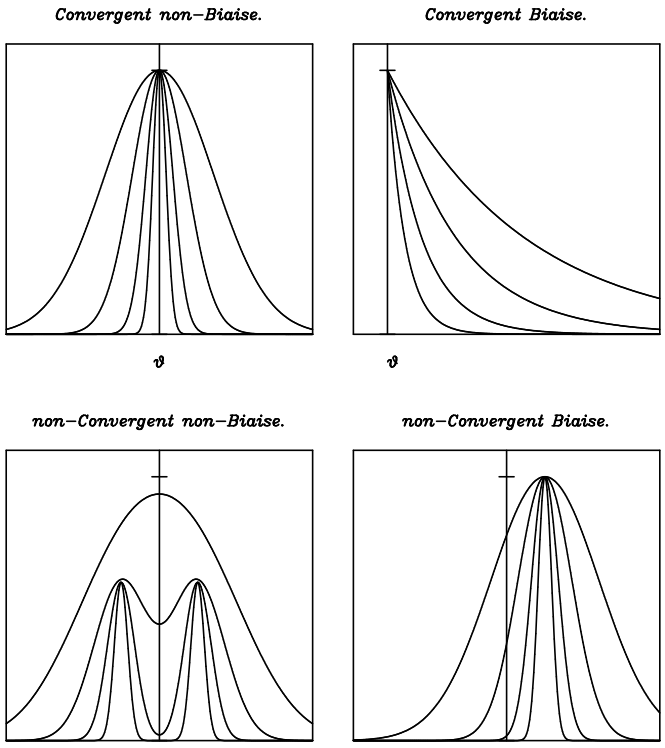


FIG. 11.1: *Illustration de l'indépendance entre convergence et absence de biais. On a représenté l'évolution des densités de probabilité de 4 estimateurs hypothétiques du paramètre θ. Les densités de probabilité ont été normalisées arbitrairement.*

11.2.2 Convergence et absence de biais.

Convergence et absence de biais sont des propriétés indépendantes. Ce fait est mis en évidence par la figure 11.1, où l'évolution de la densité de probabilité de quatre estimateurs hypothétiques $\hat{\theta}_n$ est représentée en fonction de la taille de l'échantillon n.

La propriété de convergence est plus importante que celle d'absence de biais ; il existe d'ailleurs des moyens plus ou moins faciles à mettre en œuvre pour corriger du biais, ceux-ci vont faire l'objet du chapitre suivant.

11.2.3 Les méthodes permettant de corriger du biais.

Le biais est facilement calculable.

Si l'on désire estimer de façon non-biaisée la variance σ^2 d'une population de moyenne μ inconnue, il est facile de voir que l'estimateur S^2 de σ^2 :

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - M)^2 \tag{11.10}$$

n'est pas biaisé. En effet $\mathrm{E}\{S^2\} = \mathrm{E}\{\frac{n}{n-1}S'^2\} = \frac{n}{n-1}(\sigma^2 - \sigma^2/n) = \sigma^2$. En revanche sa variance est plus élevée que celle de l'estimateur biaisé S'^2. On a en effet :

$$\mathrm{Var}(S^2) = \frac{2\sigma^4}{n-1} \quad > \quad \mathrm{Var}(S'^2) = 2\sigma^4\frac{(n-1)}{n^2} \quad . \tag{11.11}$$

En ce qui concerne l'estimation de l'écart type σ, il y a lieu de faire attention. En effet, si g est une fonction convexe, on a l'inégalité de Jensen :

$$\mathrm{E}\{g(X)\} \geq g(\mathrm{E}\{X\})\,. \tag{11.12}$$

En particulier on a : $\mathrm{E}\{S^2\} > (\mathrm{E}\{S\})^2$, et on doit alors s'attendre à ce que l'estimateur $S = \sqrt{S^2}$ soit un estimateur biaisé de σ. Dans le cas d'une variable aléatoire normale $\mathcal{N}(\mu, \sigma^2)$ par exemple, l'estimateur non-biaisé S^* de l'écart type σ est donné par l'expression :

$$S^* = k_n\Big[\frac{1}{n-1}\sum_i (X_i - M)^2\Big]^{\frac{1}{2}}, \quad k_n = \sqrt{\frac{(n-1)}{2}}\,\frac{\Gamma(\frac{n-1}{2})}{\Gamma(\frac{n}{2})}, \quad n \geq 2\,, \tag{11.13}$$

où Γ est la fonction eulérienne de 2$^{\mathrm{e}}$ espèce. On a la relation de récurrence :

$$k_2 = \sqrt{\frac{\pi}{2}}, \quad k_3 = \frac{2}{\sqrt{\pi}}, \quad k_{n+2} = \sqrt{1 - \frac{1}{n^2}}\,k_n \quad . \tag{11.14}$$

Donnons quelques valeurs de k_n :

$$k_2 \approx 1.2533, \quad k_3 \approx 1.1284, \quad k_4 \approx 1.0853, \quad k_5 \approx 1.0640\,. \tag{11.15}$$

De façon pratique, on a dès que n est assez grand :

$$S^* \approx \Big[\frac{1}{n-1.50}\sum_i (X_i - M)^2\Big]^{\frac{1}{2}} \quad . \tag{11.16}$$

Le biais n'est pas facilement calculable.

On a alors recours à des méthodes de ré-échantillonnage ou méthodes *Bootstrap*.

La méthode de Quenouille. S'il existe un biais de l'ordre de $1/n$, cette méthode permet de le réduire à l'ordre $1/n^2$. On suppose qu'il est possible de développer la valeur moyenne de l'estimateur en série entière de $1/n$. On extrait la moitié de l'échantillon, après avoir éventuellement retiré un point si l'échantillon était au départ de taille impaire. On a alors :

$$\begin{aligned} \mathrm{E}\{\hat{\theta}_n\} &= \theta + \frac{1}{n}\beta + O(\frac{1}{n^2}) \\ \mathrm{E}\{\hat{\theta}_{2n}\} &= \theta + \frac{1}{2n}\beta + O(\frac{1}{n^2}) \quad \text{d'où :} \\ \mathrm{E}\{2\widehat{\theta}_{2n} - \widehat{\theta}_n\} &= \theta + O(\frac{1}{n^2})\,. \end{aligned}$$

Le biais en $1/n$ a disparu mais, en général, la variance de ce nouvel estimateur augmentera d'un facteur de l'ordre de $1/n$. Une meilleure méthode consisterait à diviser au hasard le $2n$-échantillon en deux parts égales, évaluer les estimateurs correspondants $\hat{\theta}_n$ et $\hat{\theta}'_n$, et calculer le nouvel estimateur :

$$2\hat{\theta}_{2n} - \tfrac{1}{2}(\hat{\theta}_n + \hat{\theta}'_n) \tag{11.17}$$

La méthode du jackknife. Il existe une autre méthode qui n'augmente la variance que d'un terme en $1/n^2$, la méthode dite du « *jackknife*[1]. » Cette méthode demande plus de calculs, mais à l'heure actuelle où le calcul électronique devient de plus en plus rapide et de moins en moins cher, le jackknife est préférable à la méthode de Quenouille. Soit donc $\hat{\theta}_n$ un estimateur de θ, calculé à partir d'un n-échantillon. Développons de nouveau sa valeur moyenne en série entière de $1/n$:

$$\mathrm{E}\{\hat{\theta}_n\} = \theta + \sum_{k=1}^{\infty} \frac{a_k}{n^k} \tag{11.18}$$

On recalcule ensuite les n estimateurs $\hat{\theta}_{-i}$ en enlevant chaque fois un point i au n-échantillon. Soit $\overline{\hat{\theta}}_{n-1}$, la moyenne arithmétique de ces estimateurs. On construit finalement l'estimateur jackknife $\hat{\theta}'_n$ suivant l'expression :

$$\hat{\theta}'_n = n\hat{\theta}_n - (n-1)\overline{\hat{\theta}}_{n-1} = \hat{\theta}_n + (n-1)(\hat{\theta}_n - \overline{\hat{\theta}}_{n-1}), \tag{11.19}$$

dont l'espérance vaut :

$$\begin{aligned}
\mathrm{E}\{\hat{\theta}'_n\} &= \mathrm{E}\{n\hat{\theta}_n - (n-1)\overline{\hat{\theta}}_{n-1}\} \\
&= n\theta + n\sum_k \frac{a_k}{n^k} - (n-1)\theta - (n-1)\sum_k \frac{a_k}{(n-1)^k} \\
&= \theta + \sum_{k=1}^{\infty} \frac{a_k}{n^{k-1}} - \sum_{k=1}^{\infty} \frac{a_k}{(n-1)^{k-1}} \\
&= \theta + \sum_{k=2}^{\infty} a_k \Big[\frac{1}{n^{k-1}} - \frac{1}{(n-1)^{k-1}}\Big] \\
&= \theta - \frac{a_2}{n^2} + O(n^{-3}) \; .
\end{aligned}$$

Le biais à l'ordre $1/n$ a disparu. On montre qu'il est également possible de retirer le biais à l'ordre $1/n^2$ en considérant l'estimateur :

$$\hat{\theta}''_n = \frac{n^2\hat{\theta}'_n - (n-1)^2\,\overline{\hat{\theta}'}_{n-1}}{n^2 - (n-1)^2} \quad . \tag{11.20}$$

Mais la variance de cet estimateur augmente en général d'un terme en $1/n$.

1. Pour plus de détails voir "The jackknife–a review", Miller (1974) [50].

11.2.4 Importance des estimateurs non-biaisés.

On peut légitimement se demander pourquoi il semble si important d'obtenir des estimateurs non-biaisés, et pourquoi se donner tant de mal pour corriger d'un biais éventuel alors que des estimateurs biaisés peuvent se révéler meilleurs, dans le sens d'un moindre écart quadratique moyen, que des estimateurs non-biaisés ? Donnons un exemple qui éclairera le lecteur.

Supposons que p laboratoires collaborent, afin de déterminer la masse θ d'une nouvelle particule élémentaire. Chaque laboratoire i fournit son estimation $\hat{\theta}_n^i$, calculée à partir d'un échantillon de taille n. Supposons de plus que les différents laboratoires utilisent la même procédure expérimentale, de façon à ce que la suite des résultats $\hat{\theta}_n^1, \ldots, \hat{\theta}_n^p$ puisse être considérée comme la réalisation d'un p-échantillon i.i.d de $\hat{\theta}_n$. La population parente de cet échantillon i.i.d est la densité de probabilité de $\hat{\theta}_n$ qui, supposons-le encore, possède une moyenne μ et une variance σ^2.

Afin de prendre en compte tous ces résultats, on attribuera très probablement à θ, une valeur égale à la moyenne arithmétique sur i des $\hat{\theta}_n^i$, et on donnera une idée de la qualité de cette estimation finale de θ, en calculant l'écart type des $\hat{\theta}_n^i$. Préoccupons-nous seulement de cette moyenne arithmétique que nous notons $\langle\hat{\theta}\rangle_{n,p}$ à la manière des physiciens. C'est un estimateur non-biaisé de la moyenne μ de la population parente ; l'écart type de cet estimateur est égal à $\sigma/\sqrt{p}$. Si l'estimateur $\hat{\theta}_n$ est non-biaisé, on aura $\mu = \theta$ et donc $\mathrm{E}\{\langle\hat{\theta}\rangle_{n,p}\} = \theta$. En revanche s'il est biaisé, on aura $\mu \neq \theta$ et donc $\mathrm{E}\{\langle\hat{\theta}\rangle_{n,p}\} \neq \theta$. Cette situation, après tout, n'est pas catastrophique : il vaut peut-être mieux avoir affaire à un estimateur biaisé mais proche de θ, plutôt qu'à un estimateur non-biaisé mais de variance très grande autour de θ. Mais examinons de plus près une situation extrêmement courante.

Supposons qu'il ne soit pas possible d'augmenter la taille n de l'échantillon qui a servi à déterminer $\hat{\theta}_n^i$, parce que, par exemple, l'expérience durerait trop longtemps, deviendrait instable, ou qu'elle coûterait trop cher. Une façon habituelle de prétendre améliorer la qualité de l'estimation finale $\langle\hat{\theta}\rangle_{n,p}$, est d'augmenter p, en renouvelant l'expérience. Voyons ce qui arrive alors : $\langle\hat{\theta}\rangle_{n,p}$ est un estimateur convergent et non-biaisé de μ. Ainsi quand $p \to \infty$, d'après la loi des grands nombres, il convergera vers $\mu = \theta$ s'il est non-biaisé, mais vers $\mu \neq \theta$, s'il est biaisé ; seul l'estimateur non-biaisé sera un estimateur de θ convergent. Un estimateur de θ biaisé peut être convergent, cela n'empêche pas la moyenne arithmétique de ses réalisations de converger ailleurs que vers la valeur qu'il cherche à déterminer.

C'est parce qu'il est habituel de faire la moyenne arithmétique d'un ensemble de résultats, que l'on donne une si grande place aux estimateurs non-biaisés. Il faut cependant garder en mémoire que, si c'est la taille du n-échantillon qui peut augmenter à volonté, alors les estimateurs biaisés peuvent redevenir intéressants.

11.3 L'efficacité.

11.3.1 Ordre entre estimateurs convergents.

Il existe, la plupart du temps, plusieurs estimateurs convergents du même paramètre θ. Il semble naturel de préférer parmi tous ces estimateurs celui dont

la densité de probabilité soit la plus rapprochée de θ autrement dit, celui dont l'erreur quadratique moyenne autour de θ soit la plus petite possible. Nous chercherons donc à minimiser la quantité $\mathrm{E}\{(\hat{\theta}_n - \theta)^2\}$. Cette erreur quadratique moyenne peut s'exprimer en fonction de la variance de l'estimateur $\hat{\theta}_n$ et de son biais. En effet

$$\mathrm{E}\{(\hat{\theta}_n - \theta)^2\} = \mathrm{E}\{(\hat{\theta}_n - \mathrm{E}\{\hat{\theta}_n\})^2\} + (\mathrm{E}\{\hat{\theta}_n\} - \theta)^2 \,, \tag{11.21}$$

$$= \mathrm{Var}(\hat{\theta}_n) + b_n^2(\theta) \,. \tag{11.22}$$

Cette égalité est la version statistique du théorème de Huygens que nous avons évoqué plus haut.

Si l'estimateur est non-biaisé $b_n^2(\theta) = 0$ et $\mathrm{E}\{(\hat{\theta}_n - \theta)^2\} = \mathrm{Var}(\hat{\theta}_n)$, alors chercher l'estimateur de moindre écart quadratique moyen reviendra à chercher l'estimateur de moindre variance.

Définition 11.1. Estimateur optimal. Un estimateur non-biaisé $\hat{\theta}_{n\,\mathrm{opt}}$ sera dit optimal si, quel que soit l'estimateur $\hat{\theta}_n$, on a :

$$\mathrm{Var}(\hat{\theta}_{n\,\mathrm{opt}}) \leq \mathrm{Var}(\hat{\theta}_n) \,. \tag{11.23}$$

11.3.2 L'inégalité de Fréchet ou de Rao-Cramér.

Il est légitime de se poser la question suivante : la variance d'un estimateur peut-elle être aussi petite que l'on veut ? Pour tenter de répondre à cette question, plaçons-nous dans le cas plus général où T estun estimateur absolument correct (non-biaisé) d'une fonction τ de θ. Nous verrons plus loin que l'introduction de la fonction τ permet d'inclure les estimateurs biaisés pour θ dans la discussion.

Un tel estimateur T, non-biaisé, par définition, a pour moyenne $\mathrm{E}\{T\} = \tau(\theta)$ et pour variance $\mathrm{Var}(T) = \mathrm{E}\{(T - \tau(\theta))^2\}$. Pour calculer cette variance, on utilise l'opérateur espérance mathématique qui est un opérateur linéaire et auquel on peut associer un produit scalaire. Ce produit scalaire, en tant que tel, obéit à l'inégalité de Cauchy-Schwarz. Soit f une fonction du n-échantillon (noté $\boldsymbol{X}$) et du paramètre θ à estimer. D'après Cauchy-Schwarz on a :

$$\mathrm{E}\{(T - \tau(\theta))^2\}\ \mathrm{E}\{f(\boldsymbol{X};\theta)^2\} \geq [\mathrm{E}\{(T - \tau(\theta))\ f(\boldsymbol{X};\theta)\}]^2 \,, \tag{11.24}$$

d'où :

$$\mathrm{Var}(T) \geq \frac{\left[\mathrm{E}\{(T - \tau(\theta))\ f(\boldsymbol{X};\theta)\}\right]^2}{\mathrm{E}\{f(\boldsymbol{X};\theta)^2\}} \tag{11.25}$$

Cette inégalité devient une égalité si, et seulement si, f est proportionnelle à $T - \tau(\theta)$, c'est-à-dire si :

$$f(\boldsymbol{X};\theta) = A(\theta)\ (T - \tau(\theta)) \,. \tag{11.26}$$

Afin de trouver un bon minorant, il faut trouver un cas d'usage courant, où la fonction f soit de la forme (11.26) ci-dessus, de telle façon que la borne soit nécessairement atteinte. Etudions le cas où le n-échantillon $\boldsymbol{X}$ provient d'une loi parente normale de moyenne μ et de variance σ^2. Le paramètre à estimer θ est la moyenne μ de cette loi, la variance σ^2 est supposée connue.

Plaçons-nous dans le cas où l'échantillon $\boldsymbol{X}$ est formé à partir de variables aléatoires indépendantes et identiquement distribuées (i.i.d). On calcule alors facilement la fonction de vraisemblance d'une réalisation $\boldsymbol{x}$ de ce n-échantillon :

$$L(\boldsymbol{x}|\theta) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{x_i - \theta}{\sigma}\right)^2\right\}. \tag{11.27}$$

Eliminons le produit $\prod$ en prenant le log de l'expression :

$$\ln L = -n \ln(\sigma\sqrt{2\pi}) - \sum_{i=1}^{n} \frac{1}{2}\left(\frac{x_i - \theta}{\sigma}\right)^2 . \tag{11.28}$$

Eliminons la constante et le carré, en dérivant par rapport à θ. Il vient :

$$\frac{\partial \ln L}{\partial \theta} = \frac{1}{\sigma^2}\sum_{i=1}^{n} x_i - \frac{n\theta}{\sigma^2} = \frac{n}{\sigma^2}(m - \theta) . \tag{11.29}$$

Envisageons maintenant la fonction de vraisemblance, non pas comme une fonction dépendant d'une issue particulière $\boldsymbol{x}$ du n-échantillon, mais comme une variable aléatoire fonction de toutes les issues possibles. Notons $L(\boldsymbol{X}|\theta)$ la fonction de vraisemblance vue sous cet angle. Il vient :

$$\frac{\partial \ln L}{\partial \theta} = \frac{n}{\sigma^2}(M - \theta) . \tag{11.30}$$

On voit alors que, dans le cas particulier de l'estimation par la moyenne arithmétique M, de la moyenne μ d'une population normale de variance connue, la fonction $\partial \ln L/\partial\theta$ est précisement du type (11.26) qui transforme l'inégalité de Cauchy-Schwarz en une égalité. Dans ce cas la fonction $\tau(\theta)$ est égale à θ.

Pour une population quelconque, et pour l'estimation d'une fonction d'un paramètre θ quelconque par la statistique T, on est sûr que la variance de T sera toujours supérieure ou égale à la limite trouvée en remplaçant la fonction f par $\partial \ln L/\partial\theta$ dans l'inégalité de Cauchy-Schwarz. Effectuons donc cette opération et calculons les différentes espérances entrant dans l'équation (11.25) :

$$\mathrm{E}\left\{(T - \tau(\theta))\frac{\partial \ln L}{\partial \theta}\right\}, \qquad \text{et} \qquad \mathrm{E}\left\{\left(\frac{\partial \ln L}{\partial \theta}\right)^2\right\}, \tag{11.31}$$

c'est-à-dire en développant le premier terme :

$$\mathrm{E}\left\{\frac{\partial \ln L}{\partial \theta}\right\}; \quad \mathrm{E}\left\{T\frac{\partial \ln L}{\partial \theta}\right\} \quad \text{et} \quad \mathrm{E}\left\{\left(\frac{\partial \ln L}{\partial \theta}\right)^2\right\} . \tag{11.32}$$

Les espérances sont des intégrales et l'intégration doit se faire sur toutes les réalisations $\boldsymbol{x}$ de l'échantillon. Il vient pour le premier terme :

$$\mathrm{E}\left\{\frac{\partial \ln L}{\partial \theta}\right\} = \int \frac{\partial \ln L}{\partial \theta} L \, d\boldsymbol{x} = \int \frac{1}{L}\frac{\partial L}{\partial \theta} L \, d\boldsymbol{x} .$$

Sous réserve que l'on puisse intervertir intégration et dérivation, il en découle que le premier terme est nul. En effet :

$$\mathrm{E}\left\{\frac{\partial \ln L}{\partial \theta}\right\} = \int \frac{\partial L}{\partial \theta} \, d\boldsymbol{x} = \frac{\partial}{\partial \theta}\int L \, d\boldsymbol{x} = 0 . \tag{11.33}$$

Evaluons maintenant le deuxième terme :

$$\begin{aligned}\mathrm{E}\Big\{t\frac{\partial \ln L}{\partial \theta}\Big\} &= \int t\frac{\partial \ln L}{\partial \theta} L\, d\boldsymbol{x} = \int t\frac{\partial L}{\partial \theta}\, d\boldsymbol{x}\,,\\ &= \frac{\partial}{\partial \theta}\int tL\, d\boldsymbol{x} = \frac{\partial}{\partial \theta}\,\mathrm{E}\{T\}\,,\end{aligned}$$

et, puisque l'estimateur T est non-biaisé, $\mathrm{E}\{T\} = \tau(\theta)$, et il vient :

$$\mathrm{E}\Big\{t\frac{\partial \ln L}{\partial \theta}\Big\} = \frac{d}{d\theta}\tau(\theta) = \tau'(\theta)\,. \tag{11.34}$$

Terminons le calcul par le troisième terme :

$$\begin{aligned}\mathrm{E}\Big\{\Big(\frac{\partial \ln L}{\partial \theta}\Big)^2\Big\} &= \int \Big(\frac{\partial \ln L}{\partial \theta}\Big)^2 L\, d\boldsymbol{x} = \int \frac{\partial \ln L}{\partial \theta}\frac{\partial L}{\partial \theta}\, d\boldsymbol{x}\\ &= \frac{\partial}{\partial \theta}\int \frac{\partial \ln L}{\partial \theta} L\, d\boldsymbol{x} - \int \frac{\partial^2 \ln L}{\partial \theta^2} L\, d\boldsymbol{x}\\ &= \frac{\partial}{\partial \theta}\underbrace{\mathrm{E}\Big\{\frac{\partial \ln L}{\partial \theta}\Big\}}_{=0} - \int \frac{\partial^2 \ln L}{\partial \theta^2} L\, d\boldsymbol{x}\end{aligned}$$

soit :

$$\mathrm{E}\Big\{\Big(\frac{\partial \ln L}{\partial \theta}\Big)^2\Big\} = \mathrm{E}\Big\{-\frac{\partial^2 \ln L}{\partial \theta^2}\Big\}\,. \tag{11.35}$$

On déduit finalement de tous ces calculs l'importante inégalité de Rao-Cramér dite aussi inégalité de Fréchet :

$$\boxed{\mathrm{Var}(T) \geq \frac{[\tau'(\theta)]^2}{\mathrm{E}\Big\{-\dfrac{\partial^2 \ln L}{\partial \theta^2}\Big\}}\,.} \tag{11.36}$$

Le choix relativement arbitraire de la fonction $f = \partial \ln L/\partial\theta$ implique que cette borne inférieure n'est pas nécessairement atteinte. Pour qu'elle soit atteinte, il faut et il suffit que :

$$\frac{\partial}{\partial \theta}\ln L = A(\theta)\,(T - \tau(\theta))\,, \tag{11.37}$$

où $A(\theta)$ est une fonction quelconque du paramètre à estimer θ.

Si la variance d'un estimateur atteint cette borne, il est dit de variance minimum limite ou MVB. Si la variance d'un estimateur atteint une limite inférieure, nécessairement plus grande que la borne MVB, il est simplement dit de variance minimum ou MV, ou encore, comme nous l'avons vu, « optimal ».

Dans les calculs précédents, on a interchangé les opérateurs d'intégration et de dérivation partielle. Cela est en général possible à condition, par exemple, que les bornes d'intégration ne dépendent pas du paramètre θ. Cette condition n'est d'ailleurs pas nécessaire, si aux bornes d'intégration la densité de probabilité de la population parente est nulle, et si les dérivées premières par rapport à θ s'annulent également. Alors dans ce cas l'inégalité de Rao-Cramér reste encore valable.

11.3.3 Les estimateurs MVB.

Dans le cas où la borne MVB est atteinte on a donc :

$$\frac{\partial \ln L}{\partial \theta} = A(\theta)\,(t - \tau(\theta))\ . \tag{11.38}$$

En remplaçant directement cette expression dans l'équation (11.36) on obtient :

$$\mathrm{Var}(T) = \frac{\tau'(\theta)}{A(\theta)}\ , \tag{11.39}$$

et si $\tau(\theta) = \theta$:

$$\mathrm{Var}(T) = \frac{1}{A(\theta)}\ . \tag{11.40}$$

Ces deux dernières expressions permettent, dans certains cas, de trouver facilement la variance d'un estimateur MVB comme le montrent les exemples suivants.

▶**Exemple 11.2.** *Variance de la moyenne arithmétique d'une loi normale.* Pour un n-échantillon issu d'une loi normale de variance connue σ^2 et pour l'estimation de la moyenne θ par la moyenne arithmétique $T = M$, nous avons vu plus haut que l'on avait :

$$\frac{\partial}{\partial \theta} \ln L = \frac{n}{\sigma^2}(M - \theta) \quad \text{donc} \quad A(\theta) = \frac{n}{\sigma^2}\ ,$$
$$\text{d'où} \quad \mathrm{Var}(T) = \frac{\sigma^2}{n}\ .$$

▶**Exemple 11.3.** *Variance de la variance échantillonnée d'une loi normale.* Soit une population normale de densité de probabilité :

$$f(x) = \frac{1}{\sqrt{2\pi}\theta} \exp\Big(-\frac{x^2}{2\theta^2}\Big)\ . \tag{11.41}$$

La moyenne est connue, on peut la supposer nulle ; mais la variance θ^2 est inconnue. On a alors :

$$\frac{\partial \ln L}{\partial \theta} = \frac{n}{\theta^3}\Big(\frac{1}{n}\sum X_i^2 - \theta^2\Big) = \frac{n}{\theta^3}(S'^2 - \theta^2)\ . \tag{11.42}$$

La forme de la fonction $\partial \ln L/\partial\theta$ nous conduit naturellement à choisir comme estimateur T de θ^2, la statistique :

$$S'^2 = \frac{1}{n}\sum {X_i}^2 \tag{11.43}$$

Cet estimateur sera MVB, à la condition qu'il soit non-biaisé. On s'assure facilement de cette dernière condition car en effet :

$$\mathrm{E}\{S'^2\} = \int s^2\, L\, d\boldsymbol{x} = \frac{1}{n}\sum_i \int {X_i}^2 L\, d\boldsymbol{x} = \frac{1}{n}\sum_i \theta^2 = \theta^2\ , \tag{11.44}$$

La fonction $\tau(\theta) = \theta^2$ admet donc bien S'^2 comme estimateur MVB. On peut immédiatement calculer sa variance :

$$\mathrm{Var}(S'^2) = \frac{\tau'(\theta)}{\dfrac{n}{\theta^3}} = \frac{2\theta}{\dfrac{n}{\theta^3}} = \frac{2\theta^4}{n}\ . \tag{11.45}$$

On note habituellement σ l'écart type ici noté θ, d'où :

$$\mathrm{Var}(S'^2) = \frac{2\sigma^4}{n} . \tag{11.46}$$

On trouve ainsi directement la variance de la variance échantillonnée d'un échantillon normal de moyenne connue.

11.3.4 Efficacité et estimateur efficace.

Soit T un estimateur convergent et sans biais pour l'estimation de $\tau(\theta)$. On mesure son efficacité comme l'inverse du rapport de sa variance à la variance limite donnée par la borne MVB :

$$\mathrm{Eff}\,(T) = \frac{\mathrm{Var}_{\mathrm{MVB}}(\theta)}{\mathrm{Var}(T)} . \tag{11.47}$$

L'estimateur T sera dit *efficace* si son efficacité est de 1 (ou 100%), ou, en d'autres termes, s'il atteint sa borne MVB. Comme nous l'avons déjà fait remarquer, la borne MVB n'étant pas nécessairement atteinte, un estimateur optimal n'est pas nécessairement efficace.

11.3.5 Cas des estimateurs biaisés.

Soit T un estimateur biaisé de θ. Par définition du biais on a :

$$\mathrm{E}\,\{T\} = \theta + b(\theta) . \tag{11.48}$$

Si l'on choisit la fonction $\tau(\theta) = \theta + b(\theta)$, on a alors :

$$\mathrm{E}\,\{T\} = \tau(\theta) , \tag{11.49}$$

ce qui montre que T est un estimateur non-biaisé de $\tau(\theta)$. Appliquons la formule de Rao-Cramér pour cette fonction τ. On trouve $\tau'(\theta) = 1 + b'(\theta)$, et on a également :

$$\mathrm{Var}(T) \equiv \mathrm{E}\left\{[t - (\theta + b(\theta))]^2\right\} = \mathrm{E}\left\{(T-\theta)^2\right\} - b^2(\theta) . \tag{11.50}$$

Le terme $\mathrm{E}\left\{(T-\theta)^2\right\}$ est l'erreur quadratique moyenne de T. Nous venons de voir qu'il est toujours supérieur ou égal à la variance de T. On peut écrire finalement :

$$\mathrm{E}\left\{(T-\theta)^2\right\} = \mathrm{Var}(T) + b^2(\theta) \geq \frac{[1+b'(\theta)]^2}{\mathrm{E}\left\{-\dfrac{\partial^2 \ln L}{\partial \theta^2}\right\}} + b^2(\theta) . \tag{11.51}$$

S'il est vrai que l'erreur quadratique moyenne d'un estimateur biaisé est toujours plus grande que sa variance, l'exemple 11.4 ci-dessous montre qu'il est possible de trouver un estimateur biaisé dont l'erreur quadratique moyenne soit plus petite que la plus petite variance associée à la classe des estimateurs non-biaisés.

►**Exemple 11.4.** *Estimation du paramètre de la loi exponentielle.* Soit un n-échantillon i.i.d $(X_1, \ldots, X_n)$ issu d'une population suivant la loi exponentielle de moyenne θ :

$$f(x) = \frac{1}{\theta} \exp -\frac{x}{\theta} \quad \text{où} \quad \mathrm{E}\{X_i\} = \theta \ \text{ et } \ \mathrm{Var}(X_i) = \theta^2 . \tag{11.52}$$

Choisissons la moyenne arithmétique $\overline{X}$ comme estimateur de θ. Nous savons que c'est un estimateur non-biaisé de la moyenne de la population, et donc $\mathrm{E}\{\overline{X}\} = \theta$. Dans ce cas son erreur quadratique moyenne par rapport à θ est égale à sa variance :

$$\mathrm{E}\left\{(\overline{X} - \theta)^2\right\} = \mathrm{E}\left\{(\overline{X} - (\mathrm{E}\left\{\overline{X}\right\})^2\right\} \equiv \mathrm{Var}(\overline{X})\,. \tag{11.53}$$

Etablissons maintenant que $\overline{X}$ est MVB en calculant $\partial \ln L/\partial\theta$:

$$\ln L = -n\ln\theta - \frac{1}{\theta}\sum_{i=1}^{n} X_i\,, \tag{11.54}$$

$$\frac{\partial \ln L}{\partial\theta} = -\frac{n}{\theta} + \frac{1}{\theta^2}\sum_{i=1}^{n} X_i = \frac{n}{\theta^2}(\overline{X} - \theta)\,. \tag{11.55}$$

Ceci montre que $\overline{X}$ est bien MVB pour l'estimation de θ, d'où on tire immédiatement :

$$\mathrm{Var}(\overline{X}) = \frac{\theta^2}{n}\,. \tag{11.56}$$

En revanche l'estimateur $\hat{X} = n\overline{X}/(n+1)$ est biaisé mais son erreur quadratique moyenne est plus petite que la borne MVB θ^2/n des estimateurs non-biaisés. En effet, il est bien biaisé :

$$\mathrm{E}\{\hat{X}\} = \frac{n}{n+1}\theta, \quad b_n(\theta) = -\frac{\theta}{n+1}\,, \tag{11.57}$$

mais on a les inégalités suivantes :

$$\mathrm{Var}(\hat{X}) = \frac{n}{(n+1)^2}\theta^2 \;<\; \mathrm{E}\{(\hat{X}-\theta)^2\} = \frac{\theta^2}{n+1} \;<\; \mathrm{Var}(\overline{X}) = \frac{\theta^2}{n}\,. \tag{11.58}$$

Cet exemple montre qu'il peut être trop restrictif de se limiter à la recherche du meilleur estimateur possible dans la classe des estimateurs non-biaisés. Il peut être plus fructueux de chercher à minimiser l'erreur quadratique moyenne, c'est-à-dire la dispersion de l'estimateur autour de la valeur à estimer, plutôt que de chercher à minimiser la variance en imposant l'absence de biais. Ce sont des raisons pratiques, du type de celles évoquées plus haut, qui font que les estimateurs de moindre erreur quadratique moyenne ne sont pas très utilisés. Se limiter à la classe des estimateurs non-biaisés est aussi une simplification du point de vue mathématique.

11.3.6 L'information de Fisher.

La quantité non négative :

$$I_n(\theta) \equiv \mathrm{E}\Big\{\Big(\frac{\partial}{\partial\theta}\ln L\Big)^2\Big\} = \mathrm{E}\Big\{-\frac{\partial^2}{\partial\theta^2}\ln L\Big\}, \tag{11.59}$$

est appelée « *information de Fisher* » contenue dans le n-échantillon. La variable aléatoire $\partial \ln L/\partial\theta$ étant de moyenne nulle, l'information de Fisher n'est autre que l'inverse de la variance de cette variable aléatoire.

Pour des échantillons indépendants (i.i.d), cette grandeur ne dépend que de la taille du n-échantillon et de la densité de probabilité de la population parente. On montre facilement que $I_n(\theta) = nI_1(\theta)$. Cela signifie que dans le cas

où l'inégalité de Rao-Cramér existe, la variance d'un estimateur T décroît au plus vite comme $1/n$. L'inégalité s'écrit alors :

$$\mathrm{Var}(T) \geq \frac{(\tau'(\theta))^2}{nI_1(\theta)} \ . \tag{11.60}$$

Dans l'estimation du point milieu d'une densité de probabilité uniforme sur $]0,\theta]$, l'estimateur $\frac{1}{2}(X_{(n)} + X_{(1)})$ de la moyenne $\theta/2$ a une variance asymptotique :

$$\lim_{n\to\infty} \mathrm{Var}(\tfrac{1}{2}(X_{(n)} + X_{(1)})) = \frac{1}{2n^2} \, . \tag{11.61}$$

Cette variance asymptotique décroît plus rapidement que la limite de Rao-Cramér parce que nous sommes précisément dans un cas où cette limite ne s'applique pas. En effet, dans les calculs établissant le résultat de Rao-Cramér, l'espérance mathématique s'exprime par l'intégrale $\int_0^\theta d\boldsymbol{x}$. Dans le cas évoqué ici, le paramètre que l'on veut estimer est justement fonction de la borne supérieure de l'intégrale. On ne peut alors pas inverser l'ordre des dérivations et des intégrations et l'inégalité de Rao-Cramér ne s'applique plus.

Mentionnons pour finir ce paragraphe, que la variable aléatoire $\partial \ln L/\partial\theta$ réduite tend vers une loi normale réduite quand la taille de l'échantillon tend vers l'infini. Plus précisément on a :

$$\frac{\dfrac{\partial \ln L}{\partial\theta}}{\mathrm{E}\left\{\left(\dfrac{\partial \ln L}{\partial\theta}\right)^2\right\}} \xrightarrow{\text{loi}} \mathcal{N}(0,1) \, . \tag{11.62}$$

11.3.7 Les inégalités de Bhattacharyya.

Bhattacharyya (1946) [8, 9, 10] a généralisé l'inégalité de Fréchet Rao-Cramér. Soit T un estimateur sans biais pour $\tau(\theta)$ et L la fonction de vraisemblance. Définissons la matrice $\boldsymbol{J}$ par ses éléments J_{kl} :

$$J_{kl} = \mathrm{E}\left\{\frac{1}{L}\frac{\partial^k L}{\partial\theta^k}\frac{1}{L}\frac{\partial^l L}{\partial\theta^l}\right\} \qquad k,l = 1,\ldots,m \, , \tag{11.63}$$

et le vecteur $\boldsymbol{\tau}$ des m premières dérivées de τ, d'éléments :

$$\tau_k = \frac{\partial^k \tau(\theta)}{\partial\theta^k}, \qquad k = 1,\ldots,m \, . \tag{11.64}$$

L'inégalité d'ordre m de Bhattacharyya est donnée par l'expression suivante :

$$\mathrm{Var}(T) \geq \boldsymbol{\tau}^t \boldsymbol{J}^{-1} \boldsymbol{\tau} \ . \tag{11.65}$$

Il est clair que si $m = 1$, on retrouve l'inégalité de Fréchet Rao-Cramér et la borne MVB.

Dans le cas où la borne MVB n'est pas atteinte, les nouvelles inégalités peuvent donner, pour une certaine valeur de m, une borne inférieure plus grande que la borne MVB. Voici à titre d'exemple la borne de Bhattacharyya pour $m = 2$:

$$\mathrm{Var}(T) \geq \frac{\tau'^2}{J_{11}} + \frac{(\tau' J_{12} - \tau'' J_{11})^2}{J_{11}(J_{11}J_{22} - J_{12}^2)} \, . \tag{11.66}$$

Le premier terme est la borne MVB (de l'ordre de $1/n$), et le deuxième terme est une correction de l'ordre de $1/n^2$.

Les bornes déduites des inégalités de Rao-Cramér et de Bhattacharyya sont adaptées au cas où les erreurs de mesure sont « petites » autour du paramètre estimé. On entend par « petites » des erreurs suivant une loi proche de la loi normale. Lorsque ces erreurs sont grandes, disons proches de la loi de Cauchy, il existe alors des bornes mieux adaptées, c'est-à-dire plus grandes que les bornes de type Rao-Cramér (voir à ce sujet l'article d'Abel, 1993 [1]).

11.4 Les statistiques exhaustives.

Afin de calculer un estimateur T de $\tau(\theta)$ il n'est, dans la plupart des cas, pas nécessaire de connaître séparément chacun des éléments X_i du n-échantillon. Il suffit de connaître une ou plusieurs fonctions $t = g(X_1, \ldots, X_n)$ du n-échantillon. Il est clair que si cela est possible, on aura effectué une importante réduction des données.

Une condition nécessaire et suffisante pour que la fonction g soit une statistique « *exhaustive* » (on dit aussi « *suffisante* »), est qu'il soit possible de mettre la fonction de vraisemblance du n-échantillon sous la forme :

$$L(\boldsymbol{x}|\theta) = l(t|\theta)h(\boldsymbol{x})\,, \tag{11.67}$$

pour toutes les réalisations $\boldsymbol{x}$ et t de l'échantillon et de la statistique T.

11.4.1 Exhaustivité et information.

L'exhaustivité telle qu'elle a été définie en (11.67) a pour conséquence que la densité de probabilité conditionnelle de $\boldsymbol{x}$ (t étant donnée), ne dépend pas de θ. En effet :

$$f(\boldsymbol{x}|t) = \frac{L(\boldsymbol{x}|\theta)}{\int_{t=\mathrm{cste}} L(\boldsymbol{x}|\theta)\,d\boldsymbol{x}} = \frac{h(\boldsymbol{x})}{\int_{t=\mathrm{cste}} h(\boldsymbol{x})\,d\boldsymbol{x}}\,\cdot \tag{11.68}$$

Cela montre que la répartition des x_i sur l'hyper-surface $t = g(x_1, \ldots, x_n)$ ne dépend pas de θ (ni donc de $\tau(\theta)$), et qu'une connaissance détaillée de cette répartition n'apporterait par conséquent aucune information supplémentaire sur $\tau(\theta)$.

Nous avons employé ci-dessus le terme d'information il convient d'être plus précis et d'établir le lien qu'il existe entre exhaustivité et information.

Il est immédiat d'établir que l'information de Fisher calculée à partir de l est la même que celle calculée à partir de L. En effet, on a :

$$I_n(\theta) = \mathrm{E}\Big\{\Big(\frac{\partial \ln L}{\partial\theta}\Big)^2\Big\} = \mathrm{E}\Big\{\Big(\frac{\partial \ln l}{\partial\theta}\Big)^2\Big\}\,. \tag{11.69}$$

La réduction des données n'a donc pas fait diminuer l'information de Fisher contenue dans l'échantillon.

11.4.2 Le théorème de Fisher-Neyman.

Ce théorème montre que, réciproquement, si la probabilité conditionnelle $f(\boldsymbol{x}|t)$ de $\boldsymbol{x}$ connaissant t, ne dépend pas de θ, alors il est possible de mettre la

fonction de vraisemblance sous la forme (11.67) et que la statistique T est donc exhaustive. On peut alors choisir, pour définir l'exhaustivité, l'une ou l'autre de ces propriétés : 1) la factorisation de la fonction de vraisemblance, ou 2) l'indépendance, vis-à-vis de θ, de la probabilité conditionnnelle.

11.4.3 Statistiques exhaustives et MVB.

On a la relation :

$$\frac{\partial \ln L}{\partial \theta} = \frac{\partial \ln l(t|\theta)}{\partial \theta}\,, \tag{11.70}$$

qui exprime que $\partial \ln L/\partial\theta$ ne dépend que de t et de θ et donc que les estimateurs MVB sont à chercher dans la classe des estimateurs exhaustifs. En effet $A(\theta)(t-\tau(\theta))$ est un cas particulier du membre de droite de l'équation précédente.

On montre également que s'il existe une statistique T exhaustive pour θ, et un estimateur T_1 de θ, quelconque mais non-biaisé, alors le nouvel estimateur $p(t)$ calculé comme espérance conditionnelle de T_1 connaissant t : $p(t) = \mathrm{E}\{T_1|t\}$ est MV parmi la classe des estimateurs non-biaisés de θ. Si au départ T_1 n'était fonction que de t, c'est qu'il était déjà MV. Donc s'il existe un estimateur non-biaisé et MV de $\tau(\theta)$, il est à chercher parmi les fonctions d'une statistique exhaustive.

Seules les densités de probabilité de la forme :

$$f(x|\theta) = \exp[A(\theta)B(x) + C(x) + D(\theta)], \tag{11.71}$$

peuvent posséder une statistique exhaustive. C'est le théorème de Darmois. Cette forme englobe la plupart des densités de probabilité usuelles.

11.5 Les statistiques fiables.

Comme nous l'avons déjà vu, il existe, en général, plusieurs estimateurs du même paramètre θ. On envisage souvent, entre autres possibilités, la moyenne arithmétique, la médiane ou le point milieu de l'échantillon afin d'estimer la moyenne d'une population. Pour la loi normale par exemple, la moyenne arithmétique de l'échantillon est non-biaisée et MVB ; c'est donc le meilleur estimateur possible et il n'y a pas lieu d'en chercher un autre. Mais lequel choisir dans le cas où la population parente est de nature inconnue, ou encore dans le cas où elle est connue mais contaminée par des erreurs de mesure de nature inconnue ? Pour fixer les idées nous allons choisir trois populations parentes très différentes, et donner les variances asymptotiques des trois estimateurs précédents. Ces trois densités étant symétriques, c'est en fait la position de l'axe de symétrie que l'on cherche à déterminer. Pour simplifier nous prendrons tous les facteurs d'échelle égaux à 1. On obtient le tableau suivant, où les valeurs en caractères gras sont

les plus petites possibles :

	Loi uniforme[a]	Loi normale	Loi de Cauchy
Médiane	$\dfrac{1}{4n}$	$\dfrac{\pi}{2n}$	$\dfrac{\pi^2}{4n}$
Moyenne	$\dfrac{1}{12n}$	$\boldsymbol{\dfrac{1}{n}}$	∞
Milieu	$\boldsymbol{\dfrac{1}{2n^2}}$	$\dfrac{\pi^2}{24 \ln n}$	∞

[a] Il s'agit ici de la loi uniforme entre 0 et 1.

On voit sur cet exemple que, si l'on n'a aucune idée sur ce qu'est la population parente sous-jacente, il est plus sûr d'utiliser la médiane, afin d'obtenir une variance finie dans tous les cas. On peut considérer le tableau précédent comme la matrice des pertes d'un jeu contre la nature, où dans ce cas d'ailleurs on ne peut que perdre. On fera le choix de la médiane conformément à la stratégie minimax, c'est-à-dire celle qui minimise le maximum des pertes. La médiane est un estimateur dit *fiable (ou robuste)* pour cet ensemble de populations.

Il existe plusieurs autres estimateurs fiables de l'axe de symétrie d'une population qui suit une loi symétrique. Ils sont, suivant leur type, adaptés à un spectre plus ou moins large de populations. Donnons-en quelques uns.

- La moyenne tronquée. Considérons la moyenne tronquée bilatérale symétrique. A partir du n-échantillon ordonné $(X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)})$, on retire $2q$ points, les q plus petits et les q plus grands de façon à obtenir le nouvel échantillon $(X_{(q+1)} \leq \ldots \leq X_{(n-q)})$. Le paramètre de troncature est défini ici comme :

$$\alpha = \frac{q}{n} \,, \tag{11.72}$$

et la moyenne arithmétique du $(n-2q)$-échantillon restant sera appelée moyenne symétriquement tronquée à $100\alpha\%$ et notée T_α. On aura donc :

$$T_\alpha = \frac{1}{(n-2q)} \sum_{i=q+1}^{n-q} X_{(i)} \,. \tag{11.73}$$

Notons qu'en général on fixe α compris entre 0 et 1, et que l'on calcule ensuite $q = \lfloor \alpha n \rfloor$. Dans ce cas, la moyenne est tronquée à au plus $100\alpha\%$ et au moins $100(\alpha - 1/n)\%$.

- La moyenne Winsorisée. La procédure est la même que pour la moyenne tronquée mais on remplace les q plus petites valeurs par $X_{(q+1)}$ et les plus grandes par $X_{(n-q)}$, puis on calcule la moyenne :

$$W_\alpha = \frac{1}{n}\left[\sum_{i=q+1}^{n-q} X_{(i)} + q(X_{(q+1)} + X_{(n-q)}) \right] \tag{11.74}$$

- La médiane de Hodges-Lehmann. On fabrique le nouvel échantillon (Y_{11},-$Y_{12}, \ldots, Y_{nn}$) à partir de l'échantillon ($X_1, \ldots, X_n$) suivant la formule :

$$Y_{ij} = \tfrac{1}{2}(X_i + X_j)\,, \tag{11.75}$$

  et l'on définit la médiane de Hodges-Lehman HL comme la médiane des Y_{ij}.

- Le point milieu. C'est la statistique du point situé à égale distance des valeurs extrêmes, et qui vaut donc :

$$P = \tfrac{1}{2}[X_{(1)} + X_{(n)}]\,. \tag{11.76}$$

▶**Exemple 11.5.** *Variances asymptotiques de divers estimateurs du paramètre de localisation d'une loi de Cauchy.* Soit un n-échantillon i.i.d issu d'une loi de Cauchy de densité de probabilité :

$$f(x) = \frac{1}{\pi[1 + (x - \theta)^2]}\,. \tag{11.77}$$

Les variances asymptotiques de différents estimateurs de θ sont données par le tableau suivant :

Estimateur	Variance asymptotique
Maximum de vraisemblance	$\frac{2}{n}$
Moyenne tronquée à 38%	$\simeq \frac{2.28}{n}$
Médiane	$\frac{\pi^2}{4n} \simeq \frac{2.47}{n}$

On voit sur ce tableau que la moyenne tronquée à 38% est à peine moins efficace, mais elle est plus facile à calculer que l'estimation, dans ce cas optimale, du maximum de vraisemblance.

Les trois figures suivantes (figures 11.2, 11.3 et 11.4) illustrent graphiquement les performances de six estimateurs calculés à partir d'un échantillon de taille $n = 30$, pour 200 tirages successifs de celui-ci et cela pour trois lois différentes, la loi normale, la loi uniforme, et la loi de Cauchy. Sur ces figures, m et s désignent la moyenne et l'écart type des 200 valeurs trouvées à partir de l'estimateur considéré.

11.6 Exercices et probèmes.

Exercice 11.1. *Biais de S'^2.* Sans faire appel au théorème de Huygens calculer le biais de la variance empirique d'un échantillon (voir équation (11.9)).

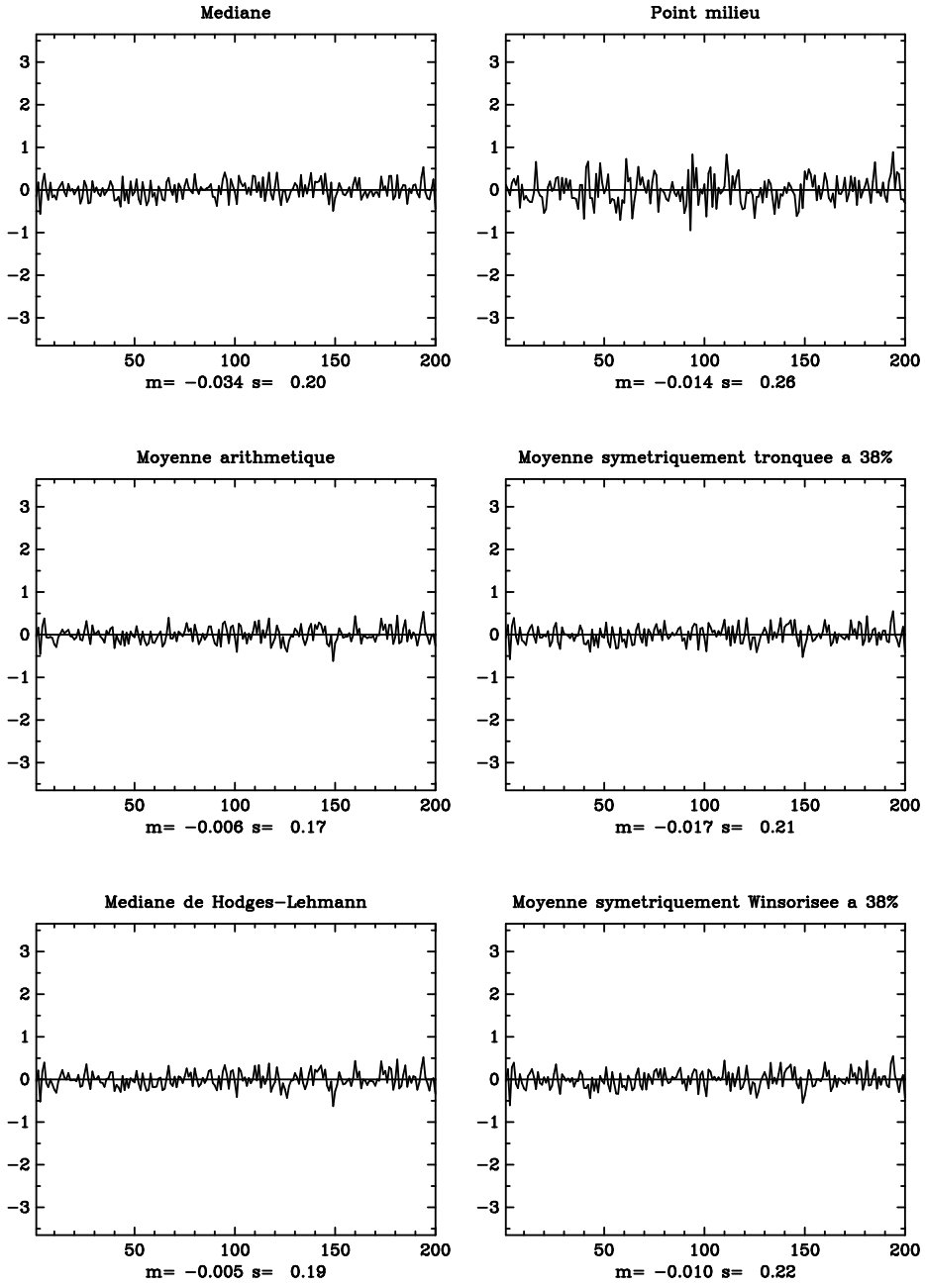


FIG. 11.2: *Performances de 6 estimateurs de la moyenne d'une loi normale. Le meilleur estimateur est la moyenne arithmétique.*

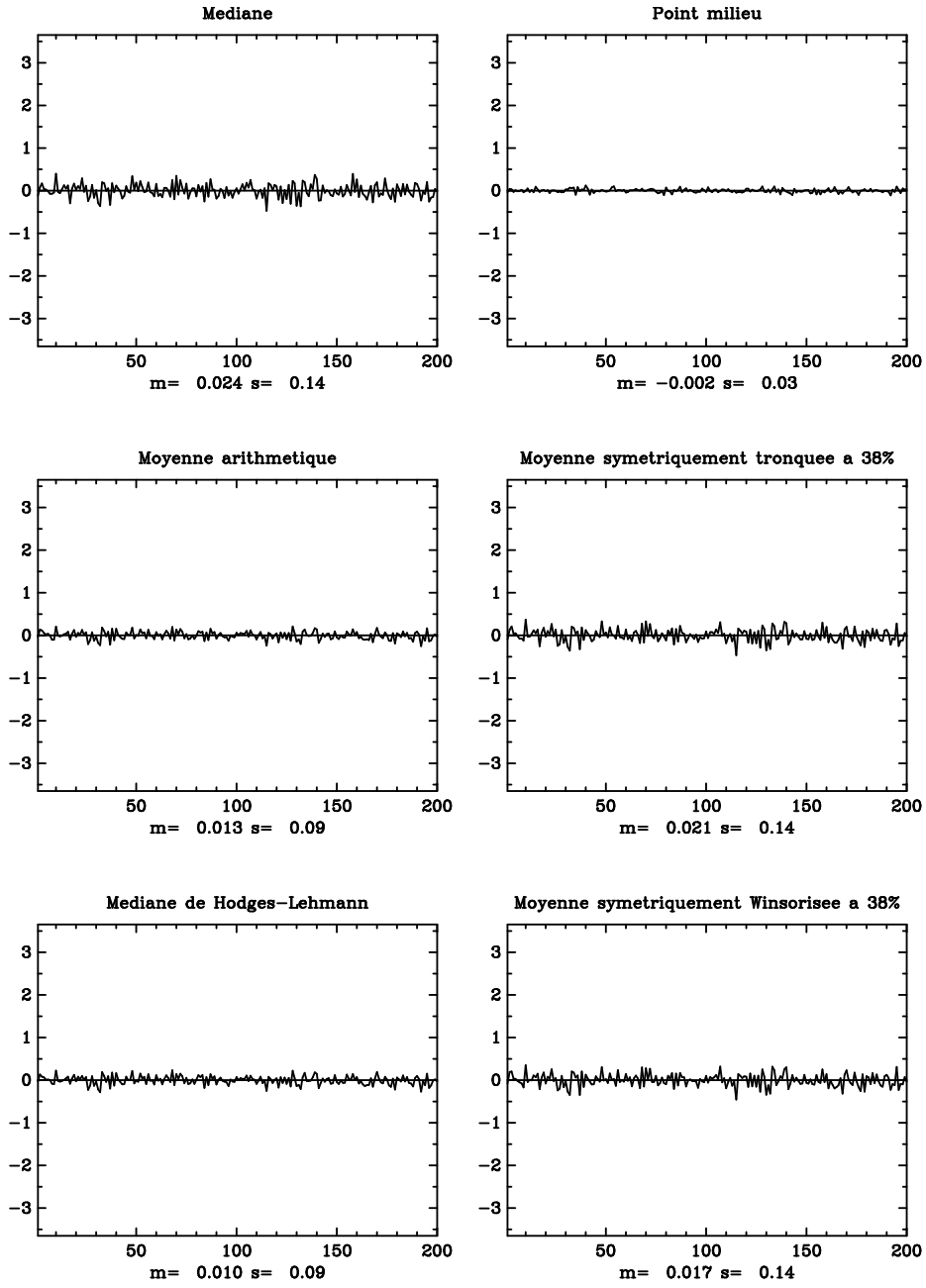


FIG. 11.3: *Performances de 6 estimateurs de la moyenne d'une loi uniforme. Le meilleur estimateur est le point milieu (la moyenne des valeurs extrêmes).*

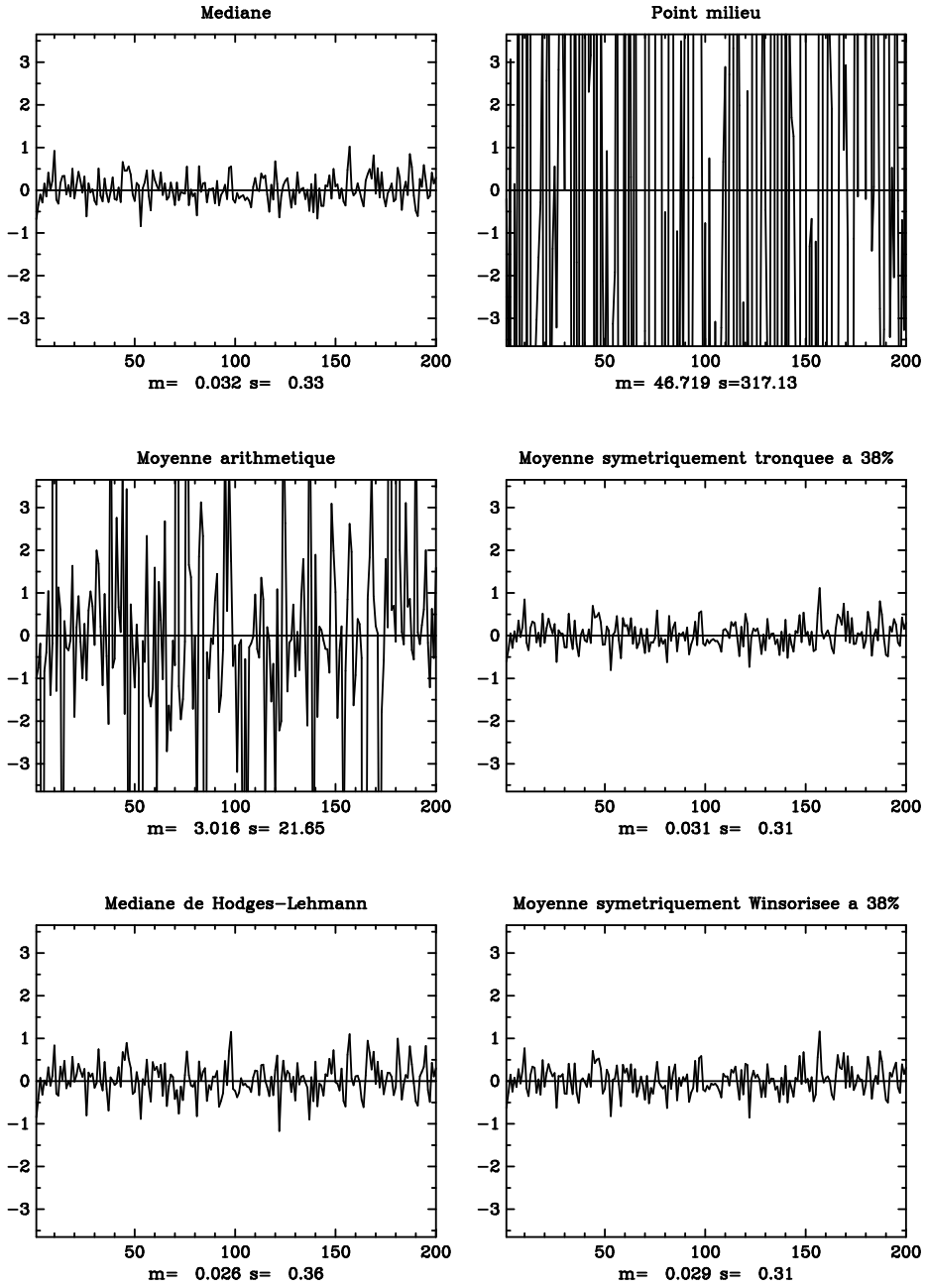


FIG. 11.4: *Performances de 6 estimateurs de la médiane d'une loi de Cauchy. Le meilleur estimateur est ici la moyenne tronquée à 38%.*

Chapitre 12

L'estimation d'intervalle.

Jusqu'à présent, à partir d'un n-échantillon, nous nous sommes attachés à essayer de trouver une estimation ponctuelle $\hat{t}$ du paramètre inconnu θ, ou d'une fonction $\tau(\theta)$ de ce paramètre inconnu. L'estimateur $\hat{t}$ étant une variable aléatoire, il est bien clair que pour différentes réalisations du n-échantillon, nous trouverons différentes valeurs issues de l'estimateur $\hat{t}$. Si cet estimateur possède des propriétés optimales, ces valeurs issues de $\hat{t}$ seront certainement proches du paramètre à estimer θ, mais elles ne nous disent pas où se trouve exactement θ. Il est naturellement impossible de savoir avec certitude où se trouve θ, mais nous pouvons tenter de le localiser, à l'aide de $\hat{t}$, en délimitant une région de l'espace du ou des paramètres où θ s'y trouverait avec la probabilité γ. Nous appellerons une telle région « *intervalle de confiance* », et la probabilité γ qui y est attachée, « *confiance* » ou « *niveau de confiance.* »

12.1 Définition de l'intervalle de confiance.

Il est en général possible de connaître, pour chaque valeur de θ, la densité de probabilité de $\hat{t}$, ou du moins sa valeur asymptotique. Soit $q(t|\theta)$ cette densité de probabilité.

▶**Exemple 12.1.** *Loi suivie par la moyenne arithmétique d'un échantillon issu d'une population normale.* Un n-échantillon $(X_1, \ldots, X_n)$ a pour population parente la loi normale $\mathcal{N}(\theta, \sigma^2)$ de moyenne inconnue θ et de variance connue σ^2. Nous avons déjà vu que la statistique :

$$\hat{t} = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X} \,, \tag{12.1}$$

suivait une loi normale de moyenne θ et de variance σ^2/n. On a donc dans ce cas $q(t|\theta) = \mathcal{N}(\theta, \sigma^2/n)$.

Pour un paramètre θ et une valeur $\gamma \in [0, 1]$ donnés, $q(t|\theta)$ étant connu, nous pouvons en principe résoudre une équation du type :

$$\Pr\left\{t_{\inf}(\theta) < \hat{t} \leq t_{\sup}(\theta)\right\} = \gamma \,. \tag{12.2}$$

C'est-à-dire qu'il est en général possible de définir une région bornée par $t_{\inf}$ et $t_{\sup}$, où les réalisations t de la variable aléatoire $\hat{t}$ s'y trouvent avec la probabilité γ. Nous supposerons que l'intervalle $t_{\inf}, t_{\sup}$ contient le paramètre θ. Il

est bien clair que l'équation (12.2) précédente admet une infinité de solutions. Parmi celles-ci on en distingue habituellement trois.

1. L'intervalle minimal.
   Afin de localiser au mieux la région où peuvent se trouver les réalisations de $\hat{t}$, il faut pour un γ donné, minimiser la quantité $t_{\sup} - t_{\inf}$. C'est cet intervalle que nous appellerons *intervalle minimal*. Il est facile de montrer que dans la limite où $\gamma \to 0$, on a $q(t_{\sup}|\theta) = q(t_{\inf}|\theta)$, et que cet l'intervalle tend vers le mode de la densité de probabilité de $\hat{t}$.

2. L'intervalle bilatéral symétrique.
   C'est celui qui, pour un γ donné, est tel que :

$$\int_{-\infty}^{t_{\inf}} q(t|\theta)\,dt = \int_{t_{\sup}}^{\infty} q(t|\theta)\,dt = \frac{1-\gamma}{2}, \tag{12.3}$$

   soit, exprimé en fonction des quantiles de $q(t|\theta)$:

$$t_{\inf} = t_{\frac{1}{2}+\frac{\gamma}{2}}\ , \quad t_{\sup} = t_{\frac{1}{2}-\frac{\gamma}{2}} \tag{12.4}$$

   Dans la limite où $\gamma \to 0$, l'intervalle tend vers la médiane de la densité de probabilité de $\hat{t}$.

3. L'intervalle central symétrique.
   On suppose que la moyenne μ de la loi existe, et que l'on peut résoudre les équations :

$$\int_{t_{\inf}}^{\mu} q(t|\theta)\,dt = \int_{\mu}^{t_{\sup}} q(t|\theta)\,dt = \frac{\gamma}{2} \tag{12.5}$$

   Dans la limite où $\gamma \to 0$, l'intervalle tend vers la moyenne de la densité de probabilité de $\hat{t}$.

Dans la suite de ce chapitre nous considérerons toujours l'intervalle bilatéral symétrique, car c'est en général celui qui est le plus facile à calculer. Nous sommes donc maintenant capables de résoudre sans ambiguïté l'équation :

$$\Pr\left\{t_{\inf}(\theta) < \hat{t} \leq t_{\sup}(\theta)\right\} = \gamma\ . \tag{12.6}$$

Mais il reste que notre problème est de donner un intervalle qui contienne θ et non $\hat{t}$ avec la probabilité γ. Supposons qu'il soit possible d'inverser les fonctions $t_{\inf}(\theta)$ et $t_{\sup}(\theta)$ pour la valeur $\hat{t}$. On aurait alors :

$$\Pr\left\{t_{\sup}^{-1}(\hat{t}) \leq \theta < t_{\inf}^{-1}(\hat{t})\right\} = \gamma\ , \tag{12.7}$$

que nous allons écrire pour plus de clarté :

$$\Pr\left\{\theta_{\inf}(\hat{t}) \leq \theta < \theta_{\sup}(\hat{t})\right\} = \gamma\ . \tag{12.8}$$

Mais il faut se garder de mal interpréter cette équation, car pour une réalisation donnée du n-échantillon :
– soit θ est dans cet intervalle et alors $\Pr\{\theta_{\inf} \leq \theta < \theta_{\sup}\} = 1$,
– soit θ n'est pas dans cet intervalle et $\Pr\{\theta_{\inf} \leq \theta < \theta_{\sup}\} = 0$.

Cela semble être en contradiction avec l'équation (12.8). Mais il faut bien comprendre que les bornes de l'intervalle $\theta_{\inf}(\hat{t})$ et $\theta_{\sup}(\hat{t})$ sont des variables aléatoires et que l'équation (12.8) n'a de sens que dans ce contexte. Cependant, dans la pratique, il faudra bien calculer des valeurs sûres $\theta_{\inf}(t)$ et $\theta_{\sup}(t)$ à partir d'une réalisation t de la variable aléatoire $\hat{t}$. Le sens à donner alors à l'intervalle de confiance (12.8) est que le paramètre θ sera compris dans l'intervalle $\theta_{\inf}(t)$, $\theta_{\sup}(t)$ avec la probabilité γ, à condition de renouveler un très grand nombre de fois l'expérience conduisant à une valeur t issue de $\hat{t}$. C'est dans le cadre d'une telle expérience de pensée, que nous parlerons d'« *intervalle de confiance* » d'un paramètre θ et du « *coefficient de confiance* » γ attaché à cet intervalle.

En d'autres termes, si l'on choisit $\gamma = 0.9$ par exemple, cela veut dire que 90% du temps, en moyenne, notre paramètre inconnu θ sera bien à l'intérieur de l'intervalle de confiance $\theta_{\inf}$,$\theta_{\sup}$ déduit à partir des réalisations successives du n-échantillon.

Illustrons cela graphiquement à l'aide de l'exemple précédent. L'estimateur $\hat{t} = \bar{X}$ de la moyenne θ d'une loi normale suit également une loi normale $\mathcal{N}(\theta, \sigma^2/n)$. Choisissons $\gamma = 0.68$, le paramètre θ étant donné. On en déduit :

$$t_{\inf}(\theta) = \theta - \frac{\sigma}{\sqrt{n}} \quad \text{et} \quad t_{\sup}(\theta) = \theta + \frac{\sigma}{\sqrt{n}} \, . \tag{12.9}$$

Portons les valeurs $t_{\inf}(\theta)$ et $t_{\sup}(\theta)$ dans le plan t, θ et répétons l'opération pour tous les θ possibles, on obtient ainsi deux droites (voir figure 12.1). Cherchons maintenant à inverser les fonctions $t_{\inf}$ et $t_{\sup}$ pour toutes les valeurs de $\hat{t}$:

$$\hat{t} = t_{\inf}(\theta_{\sup}) = \theta_{\sup} - \frac{\sigma}{\sqrt{n}} \tag{12.10}$$

$$\hat{t} = t_{\sup}(\theta_{\inf}) = \theta_{\inf} + \frac{\sigma}{\sqrt{n}} \, , \tag{12.11}$$

d'où l'on déduit l'intervalle de confiance :

$$\theta_{\inf} = \hat{t} - \frac{\sigma}{\sqrt{n}} \qquad \theta_{\sup} = \hat{t} + \frac{\sigma}{\sqrt{n}} \, \cdot \tag{12.12}$$

La vraie moyenne est égale à θ_0, nous l'ignorions, mais comme nous avons envisagé tous les θ possibles, nous avons donc construit au passage l'intervalle $t_{\inf}(\theta_0)$,$t_{\sup}(\theta_0)$ qui contient $\hat{t}$ avec la probabilité γ. Supposons qu'au cours d'une expérience réelle, nous ayons observé t_1 qui « par chance » se trouve être entre $t_{\inf}(\theta_0)$ et $t_{\sup}(\theta_0)$. Construisons l'intervalle de confiance de θ déduit à partir de ce t_1. C'est, sur la figure 12.1 l'intervalle vertical passant par la valeur t_1. Nous voyons bien que cet intervalle contient la vraie valeur θ_0 et que cet événement arrivera chaque fois que $\hat{t}$ sera dans l'intervalle $t_{\inf}(\theta_0)$, $t_{\sup}(\theta_0)$, ce qui se produit avec la probabilité γ. En revanche si par « manque de chance » nous avons observé t_2 en dehors de l'intervalle $t_{\inf}(\theta_0)$, $t_{\sup}(\theta_0)$, ce qui arrive avec la probabilité $1-\gamma$, l'intervalle de confiance ne contiendra pas θ_0 comme on le voit sur la figure 12.1, et cet événement se produira avec la probabilité $1-\gamma$. Dans l'exemple choisi on obtient finalement l'intervalle de confiance :

$$\Pr\left\{\hat{t} - \frac{\sigma}{\sqrt{n}} \leq \theta < \hat{t} + \frac{\sigma}{\sqrt{n}}\right\} = 0.68 \tag{12.13}$$

Cette interprétation géométrique est applicable chaque fois que l'on peut inverser facilement les fonctions $t_{\inf}(\theta)$ et $t_{\sup}(\theta)$.

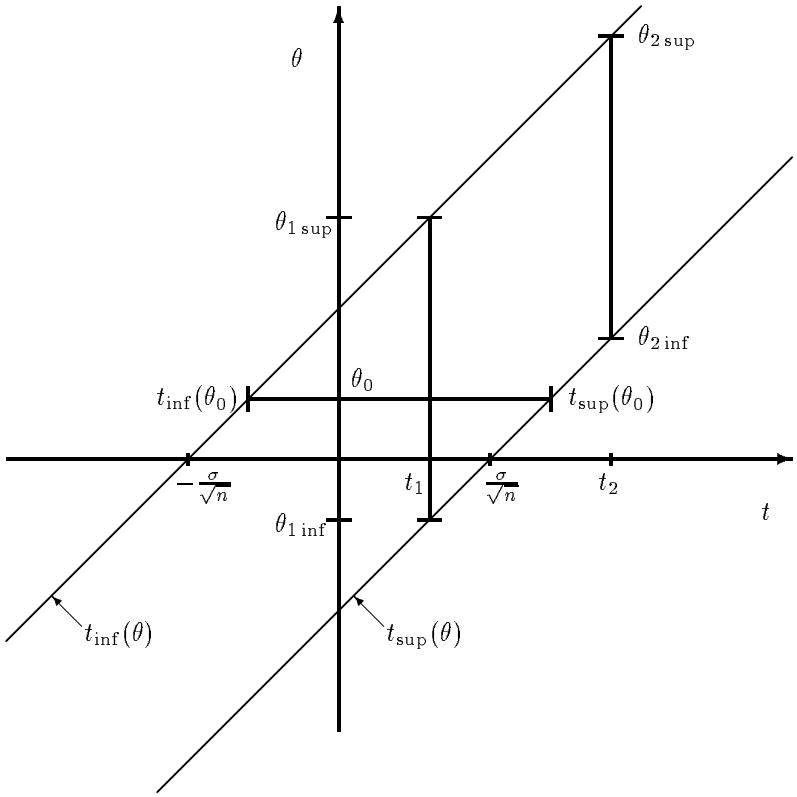


FIG. 12.1: *Construction graphique de l'intervalle de confiance de la moyenne θ_0 d'une loi normale, connaissant la variance σ^2. Sur cet exemple, l'intervalle de confiance est déterminé à partir de l'estimation t_1 ou t_2 issue d'un échantillon de taille n et pour le niveau de confiance $\gamma = 0.638$.*

12.2 Intervalle de confiance pour de grands échantillons.

Nous avons vu au cours de la démonstration de l'inégalité de Rao-Cramèr, que la fonction aléatoire $\partial \ln L/\partial\theta$, où L est la fonction de vraisemblance d'un n-échantillon, avait les caractéristiques suivantes :

$$\text{moyenne : } \mathrm{E}\Big\{\frac{\partial \ln L}{\partial \theta}\Big\} = 0$$
$$\text{variance : } \mathrm{E}\Big\{\Big(\frac{\partial \ln L}{\partial \theta}\Big)^2\Big\} = -\,\mathrm{E}\left\{\frac{\partial^2 \ln L}{\partial \theta^2}\right\}.$$

Les moyennes sont calculées sur l'espace des échantillons. Dans ces conditions, la variable aléatoire :

$$\psi = \frac{\dfrac{\partial \ln L}{\partial \theta}}{\Big[\mathrm{E}\Big\{\Big(\dfrac{\partial \ln L}{\partial \theta}\Big)^2\Big\}\Big]^{\frac{1}{2}}} \tag{12.14}$$

tend (en loi) vers la loi normale réduite $\mathcal{N}(0,1)$ lorsque la taille de l'échantillon tend vers l'infini. Il est possible de se servir de cette propriété afin de construire des intervalles de confiance, comme le montre l'exemple suivant emprunté à Kendall et Stuart (1979).

▶**Exemple 12.2.** *Intervalle de confiance du paramètre d'une loi de Poisson.* Soit X une variable aléatoire discrète $x \in \mathbb{N}$, suivant une loi de Poisson :

$$f(x;\lambda) = e^{-\lambda}\frac{\lambda^x}{x!} \tag{12.15}$$

Cherchons un estimateur de la moyenne λ à partir d'un n-échantillon $(x_1, \ldots, x_n)$. Pour cela calculons $\partial \ln L/\partial\lambda$. On obtient :

$$\frac{\partial \ln L}{\partial \lambda} = \frac{n}{\lambda}(\bar{X} - \lambda), \tag{12.16}$$

ce qui prouve que la moyenne de l'échantillon $\bar{X}$ est MVB pour l'estimation de λ, et donc que $\bar{X}$ est nécessairement une statistique exhaustive. Calculons maintenant l'information de Fisher :

$$I_n = \mathrm{E}\left\{-\frac{\partial^2 \ln L}{\partial \lambda^2}\right\}, \tag{12.17}$$

$$\frac{\partial^2 \ln L}{\partial \lambda^2} = -\frac{n\bar{X}}{\lambda^2}, \quad \mathrm{E}\left\{-\frac{\partial^2 \ln L}{\partial \lambda^2}\right\} = \frac{n}{\lambda^2}\,\mathrm{E}\left\{\bar{X}\right\}.$$

Lorsque la moyenne de la population existe, la moyenne arithmétique $\bar{X}$ de l'échantillon est toujours non-biaisée et donc :

$$I_n = \frac{n}{\lambda^2}\,\mathrm{E}\left\{\bar{X}\right\} = \frac{n}{\lambda^2}\lambda = \frac{n}{\lambda}. \tag{12.18}$$

Nous retrouvons ici le résultat bien connu où t étant un estimateur MVB de θ, alors l'information de Fisher est donnée par le terme en facteur de $(t-\theta)$ dans l'expression de $\partial \ln L/\partial\theta$, d'où ψ :

$$\psi = \frac{n}{\lambda}(\bar{X} - \lambda)\left(\frac{\lambda}{n}\right)^{\frac{1}{2}} = (\bar{X} - \lambda)\left(\frac{n}{\lambda}\right)^{\frac{1}{2}} \tag{12.19}$$

D'après ce que nous avons vu, cette quantité tend vers la loi normale réduite $\mathcal{N}(0,1)$, quand $n \to \infty$. Choisissons $\gamma = 0.95$, ce qui correspond à l'intervalle ± 1.96 pour la loi normale réduite. Il vient alors, pour une réalisation $\bar{x}$ de $\bar{X}$:

$$(\bar{x} - \lambda)\left(\frac{n}{\lambda}\right)^{\frac{1}{2}} = \pm 1.96$$
$$(\bar{x} - \lambda)^2\left(\frac{n}{\lambda}\right) = 3.84$$
$$\lambda^2 - (2\bar{x} + \frac{3,84}{n})\lambda + \bar{x}^2 = 0\,,$$

d'où l'intervalle de confiance sur la moyenne λ :

$$\lambda_{\inf} = \bar{x} + \frac{1.92}{n} + \left(\frac{3.84}{n}\bar{x} + \frac{3.69}{n^2}\right)^{\frac{1}{2}} \tag{12.20}$$

$$\lambda_{\sup} = \bar{x} + \frac{1.92}{n} - \left(\frac{3.84}{n}\bar{x} + \frac{3.69}{n^2}\right)^{\frac{1}{2}} \tag{12.21}$$

Lorsque n devient très grand on néglige les termes en $1/n$ et il reste :

$$\lambda = \bar{x} \pm \left(\frac{3.84}{n}\bar{x}\right)^{\frac{1}{2}} = \bar{x} \pm 1.96\left(\frac{\bar{x}}{n}\right)^{\frac{1}{2}}\,. \tag{12.22}$$

Mais $\mathrm{Var}(x) = \lambda \simeq \bar{x}$ que l'on notera comme d'habitude σ^2, et l'on retrouve ainsi l'intervalle de confiance de la loi normale :

$$\lambda = \bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}} \quad \text{pour} \quad \gamma = 0.95 \tag{12.23}$$

12.3 Le point de vue Bayesien.

Supposons que le paramètre inconnu soit maintenant une variable aléatoire θ, dont la densité de probabilité *a priori* $\pi(\theta)$ est connue. Nous connaissons également la densité de probabilité de l'estimateur $\hat{t}$ pour θ fixé que nous avions notée $q(t|\theta)$. En appliquant la formule de Bayes nous pouvons calculer ψ, la densité de probabilité *a posteriori* de θ, connaissant une réalisation t de l'estimateur $\hat{t}$:

$$\psi(\theta|t) = \frac{\pi(\theta)q(t|\theta)}{\int \pi(\theta)q(t|\theta)\,d\theta}\,. \tag{12.24}$$

A l'aide de cette probabilité *a posteriori* et γ étant donné on peut définir l'intervalle de confiance Bayesien comme solution de l'équation :

$$\Pr\{\theta_{\inf} < \theta \leq \theta_{\sup}\} = \int_{\theta_{\inf}}^{\theta_{\sup}} \psi(\theta|t)\,d\theta = \gamma \tag{12.25}$$

Cette équation a également une infinité de solutions et l'on choisira un des trois intervalles classiques mentionnés plus haut (voir page 224).

12.3.1 Exemple tiré de la loi normale.

Soit une réalisation $(x_1, \ldots, x_n)$ d'un n-échantillon issu d'une population normale de moyenne inconnue μ et de variance connue σ^2 égale à 1. La méthode

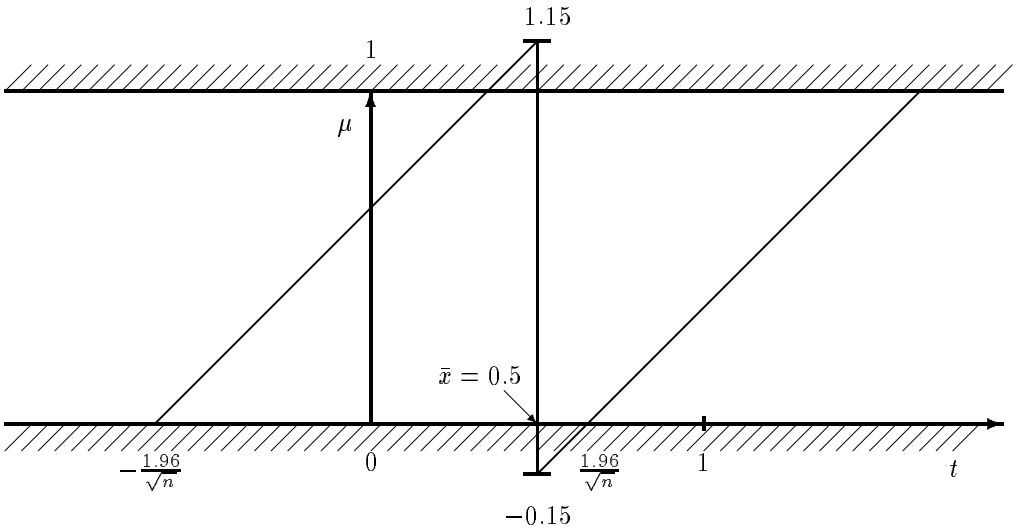


FIG. 12.2: *Extrapolation de l'intervalle de confiance sans tenir compte de l'information* a priori.

classique donne comme intervalle de confiance pour l'estimateur $\hat{t} = \bar{X}$ au niveau de confiance $\gamma = 0.95$:

$$\Pr\left\{\bar{X} - \frac{1.96}{\sqrt{n}} < \mu < \bar{X} + \frac{1.96}{\sqrt{n}}\right\} = 0.95\ . \tag{12.26}$$

Supposons maintenant que nous sachions parfaitement, pour des raisons physiques par exemple, que μ est compris entre 0 et 1, ce qui entraîne $\Pr\{0 < \mu < 1\} = 1$. La moyenne de l'échantillon $\bar{x}$ peut prendre toutes les valeurs de $-\infty$ à $+\infty$, mais supposons que nous ayons trouvé $\bar{x} = 0.5$, et que la taille de l'échantillon soit égale à $n = 9$. La méthode classique de construction de l'intervalle de confiance illustré par la figure 12.2 nous donnera alors :

$$\Pr\{-0.15 < \mu < 1.15\} = 0.95\ . \tag{12.27}$$

Cela semble être en contradiction avec les limites connues sur μ, puisque cette probabilité semblerait devoir prendre la valeur 1. En fait l'interprétation classique est toujours valable : l'intervalle aléatoire que nous construisons à partir de la variable aléatoire $\bar{X}$ contiendra bien μ dans 95% des cas si nous renouvelons indéfiniment l'expérience. Mais notre embarras devant l'interprétation de l'intervalle de confiance traduit bien le fait que nous avons délibérement ignoré l'information *a priori* sur μ, à savoir que μ est compris entre 0 et 1. Il est clair sur la figure 12.2 que nous avons abusivement extrapolé notre méthode de construction de l'intervalle de confiance, au delà de la bande interdite $[0, 1]$.

Prenons maintenant le point de vue Bayesien, en supposant qu'après tout, si μ est compris entre 0 et 1, il peut avoir une densité de probabilité *a priori*

constante dans cet intervalle, c'est-à-dire :

$$\pi(\mu) = \begin{cases} 0 & \text{si} \mu \notin [0,1]\,, \\ 1 & \text{si} \mu \in [0,1]\,. \end{cases} \tag{12.28}$$

D'autre part, $\hat{t} = \bar{x}$ suit une loi normale $\mathcal{N}(\mu, 1/n)$, d'où l'on déduit la densité de probabilité *a posteriori*:

$$\psi(\mu|\bar{x}) = \begin{cases} 0 & \text{si} \mu \notin [0,1]\,, \\ \dfrac{\exp\left[-\frac{n}{2}(\mu-\bar{x})^2\right]}{\int_0^1 \exp\left[-\frac{n}{2}(\mu-\bar{x})^2\right]\, d\mu} & \text{si} \mu \in [0,1]\,. \end{cases} \tag{12.29}$$

C'est une loi normale, centrée sur $\bar{x}$, tronquée à l'intervalle $[0,1]$ et normalisée à 1. Dans notre cas cette fonction est symétrique et les trois intervalles de confiance évoqués plus haut sont identiques. Un calcul simple nous conduit à l'intervalle de confiance Bayesien :

$$\Pr\{0.05 < \mu < 0.95\} = 0.95\,. \tag{12.30}$$

Cela est plus satisfaisant que le résultat donné par la méthode classique. Le problème de la validité du choix de la densité de probabilité *a priori* reste néanmoins en suspens.

12.4 Intervalle de confiance n-D.

12.4.1 Principe de construction.

Il s'agit maintenant de localiser un ensemble de k paramètres $\theta_1, \ldots, \theta_k$ dans l'espace $P_k \subseteq \mathbb{R}^n$ des valeurs possibles de ces paramètres, à l'aide des estimateurs $\hat{t}_1, \ldots, \hat{t}_k$. Nous désignerons par $\boldsymbol{\theta}$ l'ensemble de ces paramètres, et par $\widehat{\boldsymbol{t}}$ l'ensemble des estimateurs de ces paramètres. La démarche conduisant à définir une région de confiance dans P_k est analogue au cas 1D. Il faut d'abord commencer par calculer la densité de probabilité $q(\widehat{\boldsymbol{t}}|\boldsymbol{\theta})$. Cette opération est en principe possible si l'on suppose connu l'ensemble des $\boldsymbol{\theta}$.

Il faut délimiter ensuite dans l'espace où se répartissent les $\hat{t}_i$, une région contenant $\widehat{\boldsymbol{t}}$ avec la probabilité γ. Là aussi, il existe une infinité de façons de faire cette opération, mais la façon la plus courante consiste à définir des régions bornées par une frontière où la densité q est constante. La région ainsi définie est analogue à l'intervalle minimal du cas 1D. Sur cette frontière, il existe une relation fonctionnelle entre les $\hat{t}_i$, que l'on peut en général écrire $Q(\hat{t}_1, \ldots, \hat{t}_k) = \lambda$, où λ est une certaine constante. Une réalisation située à l'intérieur de la frontière sera telle que $Q(\hat{t}_1, \ldots, \hat{t}_k) \leq \lambda$ (ou $Q > \lambda$, mais nous supposerons que c'est le cas $< \lambda$ qui prévaut). On a alors sous cette hypothèse :

$$\Pr\{Q(\hat{t}_1, \ldots, \hat{t}_k|\boldsymbol{\theta}) \leq \lambda\} = \gamma\,. \tag{12.31}$$

Il reste maintenant à envisager la relation fonctionnelle Q, comme fonction de $\boldsymbol{\theta}$, les $\widehat{\boldsymbol{t}}$ étant connus, afin d'obtenir, dans le plan P_k des paramètres, la région de confiance maintenant définie par :

$$\Pr\{Q(\theta_1, \ldots, \theta_k|\widehat{\boldsymbol{t}}) \leq \lambda\} = \gamma\,. \tag{12.32}$$

Ce processus peut naturellement conduire à des régions de confiance pathologiques, non connexes par exemple.

12.4.2 Le cas de la loi normale 2D.

Soit un n-échantillon i.i.d formé de couples X_i, Y_i de variables aléatoires issus d'une loi normale 2D de paramètres $\mu_1, \mu_2, \sigma_1, \sigma_2$ et ρ. Supposons que l'on connaisse les paramètres σ_1, σ_2 et ρ et que l'on cherche une région de confiance pour le couple μ_1, μ_2. Le couple de variables aléatoires $\bar{X}$ et $\bar{Y}$ défini par :

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i, \tag{12.33}$$

suit une loi normale 2D (voir chapitre 10.3) de moyenne μ_1, μ_2 de matrice des variances-covariances :

$$\boldsymbol{V} = \begin{pmatrix} \dfrac{\sigma_1^2}{n} & \rho\dfrac{\sigma_1\sigma_2}{n} \\ \rho\dfrac{\sigma_1\sigma_2}{n} & \dfrac{\sigma_2^2}{n} \end{pmatrix} . \tag{12.34}$$

Il en résulte que la relation d'égale probabilité $Q(\bar{X}, \bar{Y}|\mu_1, \mu_2) = k^2$ est une ellipse d'équation :

$$Q = \frac{1}{1-\rho^2}\left[\frac{(\bar{X}-\mu_1)^2}{\sigma_1^2/n} - \frac{2\rho}{\sigma_1\sigma_2/n}(\bar{X}-\mu_1)(\bar{Y}-\mu_2) + \frac{(\bar{Y}-\mu_2)^2}{\sigma_2^2/n}\right] = k^2 , \tag{12.35}$$

qui contient la probabilité γ donnée par :

$$\gamma = 1 - e^{-\frac{1}{2}k^2} . \tag{12.36}$$

La relation Q, envisagée dans le plan μ_1, μ_2 pour une réalisation $\bar{x}, \bar{y}$ donnée, $Q(\mu_1, \mu_2|\bar{x}, \bar{y})$, est aussi une forme quadratique définissant une ellipse de confiance associée à la confiance γ. La figure 12.3 montre de telles régions de confiance.

12.5 Exemples.

12.5.1 Intervalle de confiance approximatif d'un rapport de deux variables aléatoires indépendantes et normales.

La variable aléatoire D est égale au rapport de deux variables aléatoires normales indépendantes, X suivant la loi $\mathcal{N}(\alpha, \sigma_\alpha^2)$ et Y suivant la loi $\mathcal{N}(\beta, \sigma_\beta^2)$. On définit le rapport signal sur bruit SN par la quantité

$$SN^2 = \left(\frac{\alpha}{\sigma_\alpha}\right)^2 + \left(\frac{\beta}{\sigma_\beta}\right)^2 , \tag{12.37}$$

et l'on notera $\alpha^* = \alpha/\sigma_\alpha$ et $\beta^* = \beta/\sigma_\beta$. On démontre que l'intervalle interquantile symétrique $[D_{-n}, D_{+n}]$, est donné avec une très bonne approximation par

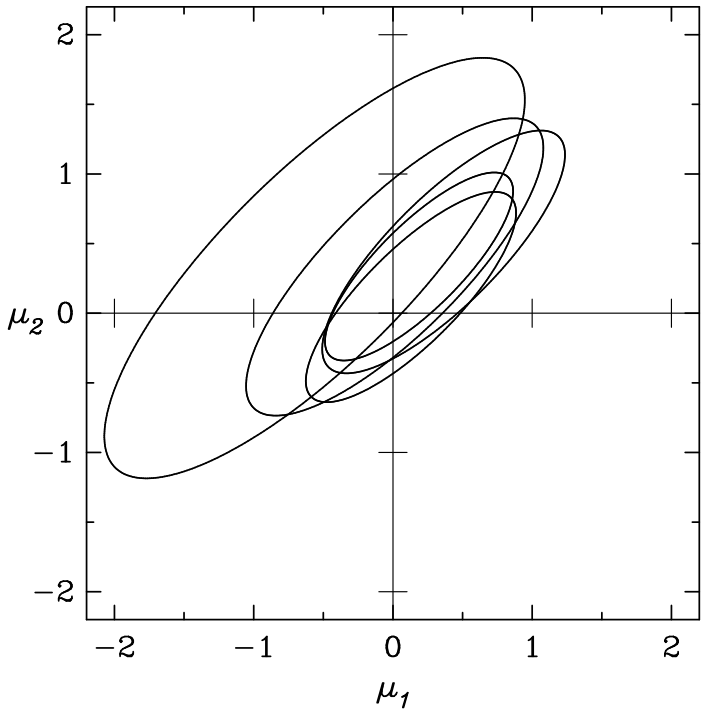


FIG. 12.3: *Région de confiance pour l'estimation simultanée de la moyenne d'une loi normale 2D, lorsque les variances et le coefficient de corrélation sont connus. Ces régions sont des ellipses calculées pour des échantillons dont la taille passe progressivement de 1 à 5. Le niveau de confiance γ est égal à 68%, et la vraie valeur de la moyenne est $\mu_1 = 0, \mu_2 = 0$. Les valeurs connues sont $\sigma_1 = \sigma_2 = 1$ et $\rho = 0.8$.*

la formule suivante

$$D_{\pm n} = \frac{\alpha \pm \dfrac{\beta^*}{(\alpha^{*2} + \beta^{*2} - n^2)^{\frac{1}{2}}} n\sigma_\alpha}{\beta \mp \dfrac{\alpha^*}{(\alpha^{*2} + \beta^{*2} - n^2)^{\frac{1}{2}}} n\sigma_\beta} \cdot \tag{12.38}$$

L'intervalle $[-n, n]$ est l'intervalle interquantile symétrique de la loi normale réduite. On a $\gamma = \Phi(n) - \Phi(-n) = 2\Phi(n) - 1$, où Φ est la fonction de répartition de la loi normale réduite. Le tableau suivant donne le contenu en probabilité γ de l'intervalle $[D_{-n}, D_{+n}]$, pour certaines valeurs de n

n	1	1.64	2	2.58	3
100γ	63.8	90.0	95.4	99.0	99.7

Cette façon de présenter les choses permet de parler de l'intervalle de confiance à "n-sigma(s)", étant entendu qu'il s'agit d'une référence à l'écart type

de la loi normale. L'approximation sur la valeur d'une borne de cet intervalle est meilleure que

$$\epsilon = 1 - \Phi(1 - n^2/SN^2)\,, \tag{12.39}$$

Posons $D_0 = \alpha/\beta$, $D_0^* = D_0\sigma_\beta/\sigma_\alpha$ et $D_{\pm n}^* = D_{\pm n}\sigma_\beta/\sigma_\alpha$. Il vient

$$D_{\pm n}^* = \frac{D_0^* \pm ((SN/n)^2 - 1)^{-\frac{1}{2}}}{1 \mp D_0^*((SN/n)^2 - 1)^{-\frac{1}{2}}}\,\cdot \tag{12.40}$$

Cette formule ne dépend que de SN/n et, quand cette quantité est donnée, l'équation précédente est celle d'une hyperbole. On ne peut pas inverser la formule précédente, car SN n'est pas uniquement fonction de D_0. Cependant, il arrive souvent que l'on ait une idée assez précise de la valeur de ce rapport signal sur bruit. Dans ces conditions, on trouve l'intervalle de confiance sur D_0

$$\Pr\{D_{0-n} \le D_0 < D_{0+n}\} = 2\Phi(n) - 1\,, \tag{12.41}$$

avec

$$D_{0\mp n} = \frac{D \pm \dfrac{\sigma_\alpha}{\sigma_\beta}((SN/n)^2 - 1)^{-\frac{1}{2}}}{1 \pm D\dfrac{\sigma_\beta}{\sigma_\alpha}((SN/n)^2 - 1)^{-\frac{1}{2}}}\,\cdot \tag{12.42}$$

où D est le rapport déduit des observations. La figure 12.4, permet de trouver les intervalles de confiance contenant D_0, pour différentes valeurs de SN/n.

Application numérique. On a mesuré les intensités de deux raies spectrales, $H_\alpha = 9.1 \pm 0.6$, $H_\beta = 2.8 \pm 0.3$. Cette mesure correspond à un rapport signal sur bruit $SN \approx 18$, on trouve en appliquant la formule (12.41), et pour le niveau de confiance $\gamma = 0.954$ correspondant à $n = 2$

$$\Pr\{2.30 \le D_0 < 5.28\} = 0.954\,. \tag{12.43}$$

12.6 Exercices.

▶**Exemple 12.3.** On procède à n expériences où un événement est susceptible de se produire avec la probabilité inconnue x. S'il se produit, on dira qu'il y a eu succès de l'expérience, dans le cas contraire, c'est un échec. On a observé p succès et q échecs, quelle est la probabilité pour que x soit compris entre les valeurs x_1 et x_2 ? (Bayes 1763 [4])

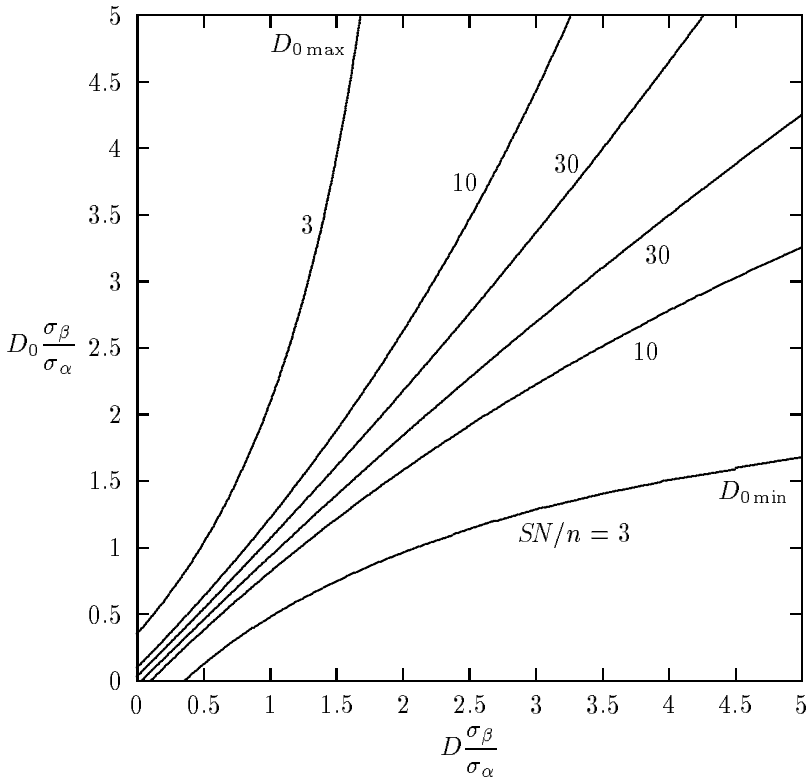


FIG. 12.4: *Abaque permettant de calculer l'intervalle de confiance d'un rapport D_0 de deux variables aléatoires normales à partir d'une mesure D de ce rapport. La quantité σ_α est l'écart type du numérateur, et σ_β l'écart type du dénominateur. Les courbes sont tracées à rapport signal sur bruit SN constant et la valeur n paramétrise le coefficient de confiance γ en unités de « sigmas » de la loi normale. On a la relation $\gamma = 2\Phi(n) - 1$, où Φ est la fonction de répartition de la loi normale réduite.*

Chapitre 13

Comment obtenir des estimateurs ?

Jusqu'à présent, nous ne nous sommes intéressés qu'aux propriétés des estimateurs, sans nous soucier de la façon pratique de les obtenir. Il existe, à cette fin, trois méthodes classiques : la méthode des moments, la méthode du maximum de vraisemblance et la méthode des moindres carrés. Nous allons exposer ici les deux premières méthodes, et consacrer tout un chapitre à la méthode des moindres carrés.

13.1 La méthode des moments.

Soit une population parente dépendant de s paramètres et ayant pour densité de probabilité :

$$f(x;\theta_1,\theta_2,\ldots,\theta_s)\,. \tag{13.1}$$

On rappelle que les moments de la population sont définis par :

$$\mu'_\nu(\theta_1,\theta_2,\ldots,\theta_s) = \mathrm{E}\left\{x^\nu\right\}\ ; \quad \text{on pose : } \mu'_1 = \mu_1 = \mu$$
$$\mu_\nu(\theta_1,\theta_2,\ldots,\theta_s) = \mathrm{E}\left\{(x-\mu)^\nu\right\}\ ; \quad \nu \geq 2\,.$$

Soit une observation $(x_1,\ldots,x_n)$ d'un n-échantillon issu de cette population. Les moments des observations sont définis par :

$$m'_\nu = \frac{1}{n}\sum_{i=1}^{n} x_i^\nu\ ; \quad \text{on pose : } m'_1 = m_1 = m \tag{13.2}$$

$$m_\nu = \frac{1}{n}\sum_{i=1}^{n}(x_i - m)^\nu\ ; \quad \nu \geq 2\,. \tag{13.3}$$

La méthode des moments consiste à résoudre le système de s équations à s inconnues obtenu en posant :

$$\mu'_\nu(\theta_1,\theta_2,\ldots,\theta_s) = m'_\nu\ ; \quad \nu = 1,\ldots,s \tag{13.4}$$

Dans ces égalités, les μ'_ν sont des scalaires alors que les m'_ν sont des réalisations des variables aléatoires M'_ν. Pour que la méthode conduise à des résultats, il faut naturellement que les μ'_ν existent. Cela n'est pas toujours le cas : la loi de Cauchy, par exemple, n'admet aucun moment à aucun ordre. Dans le cas où les μ'_ν existent, il est souhaitable que m'_ν se rapproche de μ'_ν, lorsque le nombre d'observations n augmente, et l'on demande en fait que la variable aléatoire M'_ν converge en probabilité vers la valeur μ'_ν. Ce que nous voudrions finalement, c'est que cette méthode nous conduise à trouver des estimateurs $\widehat{\theta}_{s,n}$ convergents, non-biaisés et efficaces des θ_s. Cela n'est pas toujours réalisé dans la pratique, surtout si ν est grand.

D'autre part, pour des raisons de stabilité numérique, il est préférable de résoudre le système concernant les moments centrés suivant :

$$\mu_\nu(\theta_1, \theta_2, \ldots, \theta_s) = m_\nu \; ; \quad \nu = 1, \ldots, s \, . \tag{13.5}$$

Mais dans ce cas, les estimateurs M_ν de μ_ν sont biaisés dès que $\nu \geq 2$. Nous avions déjà remarqué ce fait pour la variance. Il vaut alors mieux utiliser les observations m_2^*, m_3^* et m_4^* des estimateurs non-biaisés suivants :

$$M_2^* = \frac{n}{n-1} M_2 \, , \tag{13.6}$$

$$M_3^* = \frac{n^2}{(n-1)(n-2)} M_3 \, , \tag{13.7}$$

$$M_4^* = \frac{n(n^2-2n+3)}{(n-1)(n-2)(n-3)} M_3 - \frac{3n(2n-3)}{(n-1)(n-2)(n-3)} M_2^2 \, . \tag{13.8}$$

Enfin, pour que les variances des M_ν existent, il faut que la population parente possède des moments jusqu'à l'ordre 2ν. Dans ce cas on peut appliquer la loi des grands nombres et l'on démontre que les M_ν^* convergent en probabilité vers les μ_ν. Dans ces conditions, les estimateurs $\widehat{\theta}_{s,n}$ trouvés par la méthode des moments sont asymptotiquement corrects, de distribution asymptotiquement normale et de variance décroissant en $1/n$; mais ils sont en général peu efficaces. On les utilise habituellement comme point de départ dans la recherche d'estimateurs plus efficaces.

▶**Exemple 13.1.** *Estimation des paramètres de la loi de Laplace, par la méthode des moments.* Cherchons à estimer les paramètres μ et λ de la loi exponentielle double (ou de Laplace) de densité de probabilité :

$$f(x; \mu, \lambda) = \frac{\lambda}{2} \exp(-\lambda |x - \mu|) \, . \tag{13.9}$$

Les moments de la loi sont :

$$\mu_1 = \mu, \quad \mu_2 = \frac{2}{\lambda^2} \; . \tag{13.10}$$

Si nous cherchons à estimer μ et λ par la méthode des moments, à partir d'une observation $(x_1, \ldots, x_n)$, cette méthode nous conduit à résoudre le système formé des deux équations suivantes :

$$\mu = m_1 \quad \text{et} \quad \frac{2}{\lambda^2} = m_2^* \, . \tag{13.11}$$

Ces équations ont pour solutions :

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \text{la moyenne arithmétique des observations,}$$

$$\lambda = \left(\frac{2(n-1)}{\sum_{i=1}^{n}(x_i - m)^2}\right)^{1/2} .$$

Mais l'estimateur de μ donné par la méthode des moments n'est pas le plus efficace. La médiane de l'échantillon est un estimateur de μ deux fois plus efficace. On montre d'ailleurs qu'il est MVB.

13.2 La méthode du maximum de vraisemblance.

13.2.1 Principe de la méthode.

Soit une observation $(x_1, \ldots, x_n)$ d'un n-échantillon $(X_1, \ldots, X_n)$. La fonction de vraisemblance $L(x_1, \ldots, x_n|\theta)$ a été définie comme la densité de probabilité de l'observation, les x_i étant fixés et le paramètre θ étant considéré comme variable. Si le n-échantillon est formé de variables aléatoires indépendantes et identiquement distribuées (échantillon i.i.d), on aura :

$$L(x_1, x_2, \ldots, x_n|\theta) = f(x_1|\theta)f(x_2|\theta)\cdots f(x_n|\theta) . \tag{13.12}$$

Le principe du maximum de vraisemblance propose de choisir parmi tous les θ possibles, le paramètre $\widehat{\theta}$ qui rend la fonction de vraisemblance la plus grande possible. C'est-à-dire :

$$\forall\theta; \quad L(\boldsymbol{x}|\widehat{\theta}) \geq L(\boldsymbol{x}|\theta) . \tag{13.13}$$

En général cette équation peut se résoudre en recherchant la solution du système :

$$\frac{\partial L}{\partial\theta} = 0 \quad \text{et} \quad \frac{\partial^2 L}{\partial\theta^2} < 0 . \tag{13.14}$$

S'il y a plusieurs maxima, on choisira le plus grand. Cependant, résoudre le système (13.14) pour trouver l'estimateur du maximum de vraisemblance ne conduit pas toujours au maximum de la fonction de vraisemblance. Si, par exemple, le domaine Ω des valeurs possibles de $\boldsymbol{x}$ est borné, le maximum peut avoir lieu au bord du domaine Ω, où l'on n'aura pas nécessairement $\partial L/\partial\theta = 0$. Il faut donc se livrer à une étude critique des solutions du système (13.14), avant de déclarer que l'on a trouvé l'estimateur du maximum de vraisemblance. Dans la pratique, on cherche plutôt à maximiser le logarithme de la fonction de vraisemblance, ce qui conduit à résoudre le système :

$$\frac{\partial \ln L}{\partial\theta} = 0 \quad \text{et} \quad \frac{\partial^2 \ln L}{\partial\theta^2} < 0 . \tag{13.15}$$

13.2.2 Propriétés de l'estimateur du maximum de vraisemblance.

Donnons maintenant quelques propriétés des estimateurs du maximum de vraisemblance, que nous noterons estimateur ML. Les propriétés que nous allons mentionner ne sont valables que dans le cadre des échantillons i.i.d. Dans le cas où les populations parentes ne sont pas identiques, et même si les variables aléatoires formant l'échantillon sont indépendantes, l'utilisation des résultats qui vont suivre est sujette à caution.

Fonction d'un estimateur ML et biais.

On montre que si $\widehat{\theta}$ est un estimateur ML de θ, $\tau(\widehat{\theta})$ est également un estimateur ML de $\tau(\theta)$, ce que l'on peut exprimer par la formule :

$$\tau(\widehat{\theta}) = \widehat{\tau(\theta)} \,. \tag{13.16}$$

Cela implique que l'on doit s'attendre à ce que l'estimateur ML soit biaisé. En effet, le plus souvent $E\{\tau(\widehat{\theta})\} \neq \tau(E\{\widehat{\theta}\})$. Par exemple, si τ est une fonction convexe, on aura d'après l'inégalité de Jensen : $E\{\tau(\widehat{\theta})\} \geq \tau(E\{\widehat{\theta}\})$. Si l'estimateur ML $\widehat{\theta}$ était non-biaisé pour θ ($E\{\widehat{\theta}\} = \theta$), on aurait :

$$\mathrm{E}\{\tau(\widehat{\theta})\} \neq \tau(\mathrm{E}\{\widehat{\theta}\}) \tag{13.17}$$

$$\neq \tau(\theta) \tag{13.18}$$

et l'estimateur $\tau(\widehat{\theta})$ serait biaisé pour $\tau(\theta)$. En revanche $\tau(\widehat{\theta})$ serait non-biaisé pour $\tau(\theta)$ si τ était une fonction linéaire. La version optimiste de cette propriété est que, si l'estimateur ML est biaisé pour θ, il est peut-être possible de trouver une fonction τ, pour laquelle $\tau(\widehat{\theta})$ est un estimateur non-biaisé de $\tau(\theta)$.

Estimateurs ML et statistiques MVB.

S'il existe un estimateur t MVB de $\tau(\theta)$ et si la méthode ML donne une solution $\widehat{\theta}$, alors $t = \tau(\widehat{\theta})$, et cette solution est unique. En d'autres termes, la méthode ML fournit l'estimateur MVB si celui-ci existe.

Démonstration. S'il existe un estimateur t MVB de $\tau(\theta)$ on aura d'après l'équation (11.38) :

$$\frac{\partial \ln L}{\partial \theta} = A(\theta)(t - \tau(\theta)) \,. \tag{13.19}$$

La solution de l'équation ML doit satisfaire la condition $\partial \ln L / \partial \theta = 0$ pour $\theta = \widehat{\theta}$, d'où :

$$\left.\frac{\partial \ln L}{\partial \theta}\right|_{\theta=\widehat{\theta}} = A(\widehat{\theta})(t - \tau(\widehat{\theta})) = 0 \,. \tag{13.20}$$

La fonction $A(\widehat{\theta})$ étant en général non nulle, la seule solution de l'équation précédente, est $\tau(\widehat{\theta}) = t$. Afin de vérifier que cette solution correspond bien à

un maximum, il reste à démontrer que :

$$\left.\frac{\partial^2 \ln L}{\partial \theta^2}\right|_{\theta=\widehat{\theta}} < 0\,. \tag{13.21}$$

Or :

$$\frac{\partial^2 \ln L}{\partial \theta^2} = A'(\theta)(t-\tau(\theta)) - A(\theta)\tau'(\theta) \tag{13.22}$$

Mais, $\tau'(\theta)/A(\theta) = \mathrm{Var}(t) > 0$ et, en remplaçant θ par $\widehat{\theta}$, on obtient :

$$\left.\frac{\partial^2 \ln L}{\partial \theta^2}\right|_{\theta=\widehat{\theta}} = A'(\widehat{\theta})(t-\tau(\widehat{\theta})) - A^2(\widehat{\theta})\,\mathrm{Var}(t)\,,$$

d'où finalement :

$$\left.\frac{\partial^2 \ln L}{\partial \theta^2}\right|_{\theta=\widehat{\theta}} = -A^2(\widehat{\theta})\,\mathrm{Var}(t) < 0\,.$$

S'il existe donc un estimateur t non-biaisé et MVB pour $\tau(\theta)$, il est trouvé par la méthode du maximum de vraisemblance.

Estimateurs ML et statistiques exhaustives.

S'il existe une statistique exhaustive t pour θ, l'estimateur ML $\widehat{\theta}$ sera une fonction de t. En effet si :

$$L(\boldsymbol{x}|\theta) = l(t|\theta)h(\boldsymbol{x}) \tag{13.23}$$

le paramètre θ étant la seule variable, la solution ML $\widehat{\theta}$ qui maximise L maximise également l. Alors l'équation :

$$l(t|\widehat{\theta}) = \max_{\theta} l(t|\theta) \tag{13.24}$$

définit implicitement $\widehat{\theta}$ comme fonction de t.

De plus, si $\widehat{\theta}$ est un estimateur ML de θ et si $\tau(\theta)$ est un estimateur non-biaisé de θ, alors $\tau(\widehat{\theta})$ est un estimateur MV. Cela vient du fait que $\tau(\theta)$ ne dépend que de la statistique exhaustive t et est donc MV comme nous l'avons vu au cours de l'étude des statistiques exhaustives.

Propriétés asymptotiques.

Dans le cas général où il n'existe pas de statistique exhaustive, il n'existe que des propriétés asymptotiques. L'estimateur ML est en général asymptotiquement convergent et donc asymptotiquement non-biaisé.

13.2.3 Loi et variance de l'estimateur du maximum de vraisemblance.

Sous des conditions assez générales, on montre que l'estimateur du maximum de vraisemblance $\widehat{\theta}_n$ a une densité de probabilité $f(\widehat{\theta}_n)$ asymptotiquement

normale, de moyenne θ et de variance $I_n^{-1}(\theta)$, c'est-à-dire :

$$\lim_{n\to\infty} F(\widehat{\theta}_n) = \left[\frac{I_n(\theta)}{2\pi}\right]^{\frac{1}{2}} \int_{-\infty}^{\hat{\theta}_n} \exp\left\{-\frac{I_n(\theta)}{2}(u-\theta)^2\right\} du\,, \tag{13.25}$$

où $F(\widehat{\theta}_n)$ est la fonction de répartition de l'estimateur du maximum de vraisemblance $\widehat{\theta}_n$. Autrement dit :

$$\sqrt{I_n(\theta)}(\widehat{\theta}_n - \theta) \xrightarrow{\text{loi}} \mathcal{N}(0,1)\,. \tag{13.26}$$

Dans la mesure où la densité de probabilité de $\widehat{\theta}_n$ possède une moyenne et une variance, cela veut aussi dire que $\mathrm{E}\{\widehat{\theta}_n\} \to \theta$ et $\mathrm{Var}(\widehat{\theta}_n) \to 1/(I_n(\theta))$ et donc que l'estimateur ML est asymptotiquement non-biaisé et asymptotiquement efficace MVB. On a alors :

$$\mathrm{Var}(\widehat{\theta}_n) \simeq I_n^{-1}(\theta) = \mathrm{E}\left\{-\frac{\partial^2 \ln L(\boldsymbol{x}|\theta)}{\partial\theta^2}\right\}^{-1}, \tag{13.27}$$

et si l'échantillon est i.i.d :

$$\mathrm{Var}(\widehat{\theta}_n) \simeq n^{-1} I_1^{-1}(\theta) = n^{-1}\,\mathrm{E}\left\{-\frac{\partial^2 \ln f(x|\theta)}{\partial\theta^2}\right\}^{-1}. \tag{13.28}$$

Pour calculer la variance asymptotique de l'estimateur ML $\widehat{\theta}_n$ il semble qu'il soit nécessaire de calculer une moyenne. En fait, ce calcul se simplifie considérablement dans le cas limite où $n \to \infty$ et où les estimateurs $\widehat{\theta}$ tendent à devenir MVB. On montre, en effet, qu'il suffit de ne connaître qu'une valeur de l'expression dont on fait la moyenne. Etablissons ce fait pour un estimateur MVB t de $\tau(\theta)$. La variance de cet estimateur t (MVB) est donnée par la formule de Rao-Cramér appliquée au cas MVB et donnée par l'équation (11.39) :

$$\mathrm{Var}(t) = \frac{\tau'(\theta)}{A(\theta)}\,. \tag{13.29}$$

Par ailleurs, on a déjà vu en (13.22), que :

$$\frac{\partial^2 \ln L}{\partial\theta^2} = A'(\theta)(t - \tau(\theta)) - A(\theta)\tau'(\theta)\,. \tag{13.30}$$

On peut remplacer t par $\tau(\widehat{\theta})$, car si $\widehat{\theta}$ est l'estimateur fourni par la méthode ML, $\tau(\widehat{\theta})$ est également, d'après (13.16) l'estimateur ML de $\tau(\theta)$. De plus si l'estimateur t est MVB, il est trouvé par la méthode ML et donc $t = \tau(\widehat{\theta})$. Il vient donc :

$$\frac{\partial^2 \ln L}{\partial\theta^2} = A'(\theta)(\tau(\widehat{\theta}) - \tau(\theta)) - A(\theta)\tau'(\theta)\,, \tag{13.31}$$

et :

$$\left.\frac{\partial^2 \ln L}{\partial\theta^2}\right|_{\widehat{\theta}=\theta} = -A(\theta)\tau'(\theta)\,. \tag{13.32}$$

Exprimons ce résultat en fonction de la variance de $t = \tau(\widehat{\theta})$ donnée en (13.29). On trouve finalement :

$$\mathrm{Var}(\tau(\widehat{\theta})) = -(\tau'(\theta))^2 \left(\frac{\partial^2 \ln L}{\partial \theta^2}\bigg|_{\widehat{\theta}=\theta} \right)^{-1} . \tag{13.33}$$

Il n'est plus nécessaire de calculer la moyenne de la variable aléatoire $-\partial^2 \ln L/\partial\theta^2$, il suffit simplement de l'évaluer en un point.

Maintenant, si $\tau(\widehat{\theta})$ n'est pas MVB, mais possède une variance finie quand $n \to \infty$, alors, d'après ce que nous avons vu plus haut, il est asymptotiquement MVB. On aura donc de façon asymptotique :

$$\mathrm{Var}(\tau(\widehat{\theta})) \simeq -(\tau'(\theta))^2 \left(\frac{\partial^2 \ln L}{\partial \theta^2}\bigg|_{\widehat{\theta}=\theta} \right)^{-1} . \tag{13.34}$$

En prenant la fonction particulière $\tau(\theta) = \theta$ on a aussi :

$$\mathrm{Var}(\widehat{\theta}) \simeq -\left(\frac{\partial^2 \ln L}{\partial \theta^2}\bigg|_{\widehat{\theta}=\theta} \right)^{-1} . \tag{13.35}$$

Illustrons ce résultat à l'aide d'un exemple emprunté à Kendall et Stuart [41].

▶**Exemple 13.2.** *Estimation ML de l'écart type d'une population normale de moyenne connue.* Nous supposerons que la moyenne μ est nulle. Avec cette convention la densité de la population est donnée par :

$$f(x) = \frac{1}{\sqrt{2\pi}\theta} \exp\left(-\frac{x^2}{2\theta^2}\right) . \tag{13.36}$$

Soit un échantillon $(x_1, \ldots, x_n)$ issu de cette population. Calculons le logarithme neperien et la dérivée de la fonction de vraisemblance de cet échantillon :

$$\ln L(\boldsymbol{x}|\theta) = -\frac{n}{2}\ln(2\pi) - n\ln\theta - \frac{1}{2\theta^2}\sum_{i=1}^{n} x_i^2 , \tag{13.37}$$

$$\frac{\partial \ln L}{\partial \theta} = -\frac{n}{\theta} + \frac{1}{\theta^3}\sum_{i=1}^{n} x_i^2 = \frac{n}{\theta^3}\left(\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \theta^2\right) . \tag{13.38}$$

On sait que la statistique $\widehat{\theta} = (\frac{1}{n}\sum x_i^2)^{\frac{1}{2}}$ est un estimateur convergent de θ (voir chapitre 9.5.2, page 183). L'équation (13.37) montre que $\widehat{\theta}$ est une statistique exhaustive, mais (13.38) montre qu'elle n'est pas MVB. En revanche, en tant que statistique exhaustive, c'est aussi l'estimateur du maximum de vraisemblance. Calculons sa variance, et pour cela, conformément à l'équation (13.35), calculons la dérivée seconde de $\ln L$:

$$\frac{\partial^2 \ln L}{\partial \theta^2} = \frac{n}{\theta^2}\left(1 - \frac{3\widehat{\theta}^2}{\theta^2}\right) . \tag{13.39}$$

Remplaçons maintenant $\widehat{\theta}$ par θ, comme il est prescrit par l'équation (13.35). Il vient :

$$\mathrm{Var}(\widehat{\theta}) \simeq \frac{\theta^2}{2n} . \tag{13.40}$$

L'écart type σ, d'une population normale de loi $\mathcal{N}(\mu, \sigma^2)$ a donc pour estimateur ML :

$$\widehat{\sigma} = \left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2\right]^{\frac{1}{2}}, \tag{13.41}$$

et sa variance asymptotique vaut :

$$\mathrm{Var}(\widehat{\sigma}) = \frac{\sigma^2}{2n}\,. \tag{13.42}$$

Donnons maintenant une formule encore plus approximative, mais très utile dans la pratique. Posons $\Delta\widehat{\theta}$ égal à l'écart type de $\widehat{\theta}$. On a alors avec cette notation, $\mathrm{Var}(\widehat{\theta}) = \Delta\widehat{\theta}^2$. Posons de plus, pour alléger l'écriture, $h(\theta) = \ln L(\boldsymbol{x}|\theta)$, et effectuons un développement limité de h autour de $\widehat{\theta}$. Il vient :

$$h(\widehat{\theta} + \Delta\theta) = h(\widehat{\theta}) + h'(\widehat{\theta})\Delta\theta + \tfrac{1}{2}h''(\widehat{\theta})(\Delta\theta)^2 + O(\Delta\theta)^3\,. \tag{13.43}$$

Mais, $h'(\widehat{\theta}) = 0$, et d'après (13.35), $h''(\widehat{\theta}) \simeq -\,\mathrm{Var}(\widehat{\theta})^{-1} = -(\Delta\theta)^{-2}$, d'où au deuxième ordre en θ :

$$h(\widehat{\theta} + \Delta\theta) \approx h(\widehat{\theta}) - \tfrac{1}{2}\,. \tag{13.44}$$

On trouve alors l'écart type $\Delta\theta$ de l'estimation du maximum de vraisemblance $\widehat{\theta}$ comme solution de l'équation :

$$\ln L(\boldsymbol{x}|\widehat{\theta} + \Delta\theta) = \ln L(\boldsymbol{x}|\widehat{\theta}) - \tfrac{1}{2}\,. \tag{13.45}$$

On peut dire aussi, que cet écart type $\Delta\theta$ est trouvé par l'intersection du graphe de la fonction de vraisemblance, avec une droite horizontale au niveau de l'ordonnée du maximum moins $\frac{1}{2}$.

Pour calculer pratiquement un estimateur ML, on est donc amené à maximiser la fonction $\ln L(\theta)$. Les méthodes les plus performantes pour rechercher le maximum d'une fonction, d'une ou plusieurs variables, sont les méthodes dites « quasi-Newton. » Ces méthodes reposent sur le calcul de la dérivée seconde de la fonction dont on cherche le maximum. Si l'on emploie une telle méthode, on aura automatiquement la valeur de $(\ln L)''$ au maximum $\widehat{\theta}$. L'inverse de cette quantité, changé de signe, nous donnera alors une approximation de la variance asymptotique de l'estimation. Ce faisant, on a remplacé θ par $\widehat{\theta}$ dans l'expression (13.35), et il s'agit alors d'une approximation de la variance asymptotique.

Dans le cas où l'on cherche à estimer conjointement plusieurs paramètres $(\theta_1, \ldots, \theta_s)$, on doit chercher le maximum de la fonction de vraisemblance $\ln L(\boldsymbol{x}|\theta_1, \ldots, \theta_s)$. La matrice des variances-covariances de l'estimateur $(\widehat{\theta}_1, \ldots, \widehat{\theta}_s)$ est donnée asymptotiquement par l'inverse de la matrice des dérivées secondes de $\ln L$, ou comme on dit du « *Hessien* » de $\ln L$. Cela peut s'exprimer symboliquement par la relation :

$$[\boldsymbol{V}_{\widehat{\boldsymbol{\theta}}}]_{ij} = \left(\left. \frac{\partial^2 \ln L}{\partial\theta_i \partial\theta_j} \right|_{\widehat{\theta}=\theta} \right)^{-1}\,. \tag{13.46}$$

On pourra également se contenter d'une expression approchée de $\boldsymbol{V}_{\widehat{\theta}}$ en posant $\theta = \widehat{\theta}$ dans l'expression précédente.

13.3 Exemples.

13.3.1 Estimation d'un rapport : le décrément de Balmer.

Le décrément de Balmer D_0 est égal au rapport des intensités des raies H_α et H_β de l'hydrogène atomique. Afin d'en déterminer la valeur au centre

d'une galaxie active, on a observé le spectre d'émission de l'hydrogène au cours de 4 observations du centre de cette galaxie, étalées sur environ un an. Les observations sont données par le tableau 13.1. Les intensités de H_α et de H_β

	H_α	H_β	H_α/H_β
4-dec-80	2.8 ± 0.3	0.8 ± 0.2	3.5
10-jan-81	2.4 ± 0.3	0.5 ± 0.2	4.8
20-sep-81	4.2 ± 0.3	1.7 ± 0.2	2.5
11-jan-82	9.1 ± 0.6	2.8 ± 0.3	3.3

TAB. 13.1: *Quatre observations de l'intensité des raies d'émission H_α et H_β de l'hydrogène atomique. Le Décrément de Balmer D_i est égal au rapport: H_α/H_β, de l'intensité de ces raies.*

ont varié pendant cette période, mais en supposant que le décrément de Balmer lui, n'a pas varié, on demande qu'elle est l'estimation du maximum de vraisemblance de D_0 ?

Introduisons maintenant nos notations. On notera α_i et β_i les observations de l'intensité des raies H_α et H_β à l'époque t_i ; à ces observations correspond la valeur D_i du décrément de Balmer. Les valeurs moyennes de la loi suivie par les α_i et les β_i, seront notées μ_{α_i} et μ_{β_i}, et, par hypothèse, le décrément de Balmer cherché $D_0 = \mu_{\alpha_i}/\mu_{\beta_i}$ ne dépend pas de t_i. Nous supposerons que les erreurs de mesure sur les intensités des raies suivent une loi normale, ne sont pas corrélées, et ont pour écart type σ_{α_i} et σ_{β_i}. Les écart types sont connus et leur valeurs sont données par le tableau 13.1. Avec ces notations, les α_i et β_i sont des variables aléatoires normales respectivement égales à : $\mathcal{N}(\mu_{\alpha_i}, \sigma^2_{\alpha_i})$ et $\mathcal{N}(\mu_{\beta_i}, \sigma^2_{\beta_i})$.

Les formules (6.105) et (6.106) nous donnent la densité de probabilité f du rapport D_i, en fonction des paramètres $\mu_{\alpha_i}, \mu_{\beta_i}, \sigma^2_{\alpha_i}, \sigma^2_{\beta_i}$. De ces quatre paramètres, seuls σ_{α_i} et σ_{β_i} sont connus, et on remarquera que la fonction f dépend de μ_{α_i} et μ_{β_i} séparément et non par le seul intermédiaire de leur rapport. Cela nous interdit de considérer f comme la fonction de vraisemblance de D_i. En d'autres termes, D_i n'est pas une statistique exhaustive.

Il est utile, pour saisir toute l'étendue du problème, de tracer la densité de probabilité f en supposant, ce qui est naturellement faux, que $\mu_{\alpha_i} = \alpha_i$ et $\mu_{\beta_i} = \beta_i$, et c'est ce qui a été fait sur la figure 13.1. Il est manifeste sur cette figure que le mode de cette densité est systématiquement plus petit que la valeur $D_0 = \mu_{\alpha_i}/\mu_{\beta_i}$, ce qui veut dire que, le plus souvent, le rapport observé $D_i = \alpha_i/\beta_i$ sera plus petit que le vrai rapport D_0. On sait par ailleurs que cette densité ne possède pas de moyenne, et il est donc catastrophique d'estimer un rapport par la moyenne arithmétique des rapports déduits des observations.

Ces remarques étant faites, revenons à notre problème d'estimation. Comme il n'est pas possible de passer par l'intermédiaire de D_i, il faut estimer les μ_{α_i} et μ_{β_i} séparément. Dans le cas qui nous intéresse, on a la relation $\mu_{\alpha_i} = D_0\mu_{\beta_i}$, ce qui fait que l'on doit estimer cinq paramètres (les μ_{α_i} ou les μ_{β_i} et D_0), à l'aide des huit observations données par le tableau 13.1. Les lois suivies par les α_i et par les β_i étant normales et indépendantes, on trouve immédiatement le log de la fonction de vraisemblance des observations, à la constante additive

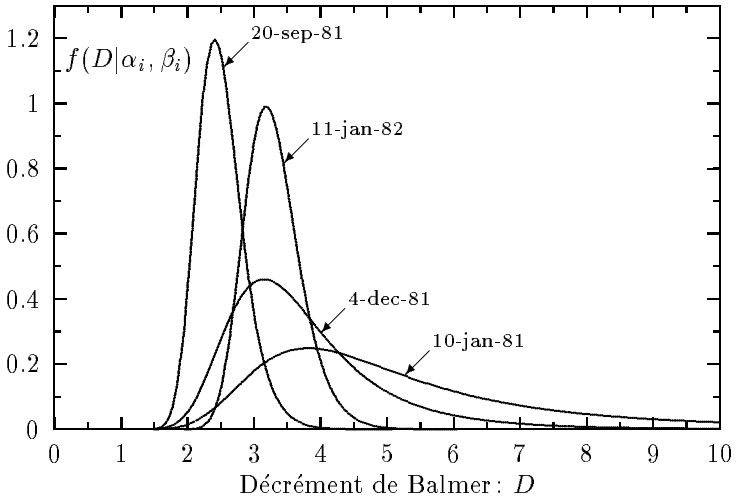


FIG. 13.1: *Ce graphe représente les densités de probabilité du rapport de l'intensité des raies de l'hydrogène H_α/H_β pour quatre observations indépendantes. Les densités de probabilité sont identifiées par la date de l'observation correspondante.*

$-\sum_{i=1}^{4} \ln(2\pi\sigma_{\alpha_i}\sigma_{\beta_i})$ près que l'on peut négliger, soit :

$$\ln L = -\frac{1}{2}\left(\sum_{i=1}^{4} \frac{(\alpha_i - \mu_{\alpha_i})^2}{\sigma_{\alpha_i}^2} + \frac{(\beta_i - \mu_{\beta_i})^2}{\sigma_{\beta_i}^2}\right) , \tag{13.47}$$

les μ_{α_i} et μ_{β_i}, étant soumis aux quatre contraintes F_i suivantes :

$$F_i(\mu_{\alpha_i}, \mu_{\beta_i}) = \mu_{\alpha_i} - D_0\mu_{\beta_i} = 0\,; \quad i = 1, \ldots, 4\,. \tag{13.48}$$

Cherchons le maximum par la méthode des multiplicateurs de Lagrange. Soient λ_i les quatre multiplicateurs, correspondant aux quatre contraintes. Les estimateurs cherchés sont solutions du système :

$$\frac{\partial}{\partial\mu_{\alpha_i}}(\ln L - \lambda_i F_i) = \frac{\alpha_i - \widehat{\mu}_{\alpha_i}}{\sigma_{\alpha_i}^2} - \lambda_i = 0 \tag{13.49a}$$

$$\frac{\partial}{\partial\mu_{\beta_i}}(\ln L - \lambda_i F_i) = \frac{\beta_i - \widehat{\mu}_{\beta_i}}{\sigma_{\beta_i}^2} - \lambda_i \widehat{D}_0 = 0 \tag{13.49b}$$

$$\frac{\partial}{\partial D_0}(\ln L - \sum_{i=1}^{4} \lambda_i F_i) = \sum_{i=1}^{4} \lambda_i \widehat{\mu}_{\beta_i} = 0\,. \tag{13.49c}$$

Nous ne donnerons pas la solution complète du système, nous nous contenterons seulement de l'estimation $\widehat{D_0}$. Il vient :

$$\widehat{\mu}_{\alpha_i} = \widehat{D}_0\widehat{\mu}_{\beta_i} \tag{13.50}$$

$$\widehat{\mu}_{\beta_i} = \frac{\widehat{D}_0\sigma_{\beta_i}^2\alpha_i + \sigma_{\alpha_i}^2\beta_i}{\sigma_{\alpha_i}^2 + D_0^2\sigma_{\beta_i}^2} \tag{13.51}$$

$$\lambda_i = \frac{\alpha_i - \widehat{D}_0\beta_i}{\sigma_{\alpha_i}^2 + D_0^2\sigma_{\beta_i}^2}\,, \tag{13.52}$$

ce qui montre que l'estimation des paramètres μ_{α_i} et μ_{β_i}, ne dépend que de l'estimation de D_0. Remplaçons ces valeurs dans l'expression de $\ln L$. Il vient :

$$\ln L(\widehat{D}_0|\boldsymbol{\alpha},\boldsymbol{\beta}) = \max_D -\frac{1}{2}\sum_{i=1}^{4} \lambda_i^2(\sigma_{\alpha_i}^2 + D^2\sigma_{\beta_i}^2) \tag{13.53}$$

$$= \max_D -\frac{1}{2}\sum_{i=1}^{4} \frac{(\alpha_i - D\beta_i)^2}{\sigma_{\alpha_i}^2 + D^2\sigma_{\beta_i}^2}\,. \tag{13.54}$$

La figure 13.2 représente le graphe de cette fonction. L'estimation du maximum de vraisemblance donne $\widehat{D}_0 = 3.07$. L'erreur sur cette valeur, erreur due à la présence de bruit dans les observations, et calculée grâce à la formule (13.35), est de 0.38. L'estimateur du maximum de vraisemblance de D_0 est donc :

$$\widehat{D}_0 = 3.07 \pm 0.38\,. \tag{13.55}$$

Faisons, à propos de cet exemple, plusieurs remarques.

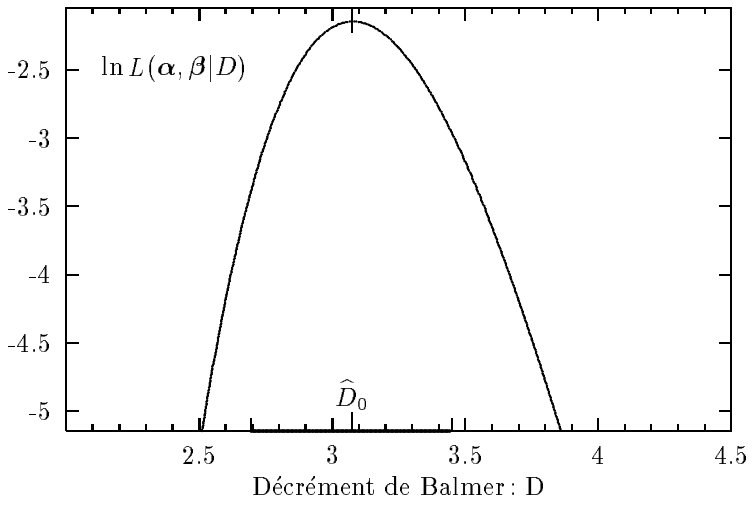


FIG. 13.2: *Estimation du décrément de Balmer du gaz hydrogène présent au centre d'une galaxie active variable. La courbe représente le logarithme de la fonction de vraisemblance de 4 observations. Le vrai décrément D_0 est inconnu mais il est supposé être constant. L'estimation donnée par la méthode du maximum de vraisemblance $\widehat{D}_0$ vaut* 3.07. *L'écart type asymptotique de $\widehat{D}_0$, estimé à partir de la courbure au sommet de la courbe, $\sigma_{\widehat{D}_0} = (-\partial^2 \log L/\partial D^2)^{-1/2}$ vaut* 0.38. *Notre estimation de D_0 est donc* 3.07 ± 0.38. *Le domaine couvert par ces valeurs est représenté en gras sur l'axe horizontal.*

1. La variance de l'estimation du ML est en fait l'inverse de la courbure de la fonction $\ln L$, au maximum de celle-ci. Plus cette fonction est « piquée » sur l'estimation, plus la variance de l'estimation sera petite, ce qui est conforme à l'intuition.

2. On a utilisé une formule asymptotique pour calculer la variance de l'estimation. C'est une approximation dont il faut aussi avoir conscience.

En fait, dans les conditions physiques normales qu'on peut s'attendre à trouver au centre d'une galaxie, le décrément de Balmer doit avoir une valeur comprise entre 2.7 et 2.9 On voit que notre estimation 3.07 ± 0.38, n'est pas contradictoire avec l'hypothèse que le gaz présent au centre de la galaxie observée reste toujours dans des conditions normales, même au cours d'une phase active où l'intensité de l'émission de l'hydrogène varie. Il faut bien insister sur le peu de crédit qu'il faut accorder à un rapport de deux mesures, sans une analyse critique des erreurs. Ayant mesuré des rapports aussi différents que 2.5 et 4.8, on aurait pu en déduire que le décrément de Balmer avait varié de façon significative d'une observation à l'autre. En fait, il n'en était probablement rien, et le décrément de Balmer avait très certainement une valeur toujours proche de 2.9. Le problème de savoir s'il était raisonnable de penser que ce rapport était constant, aurait pu être étudié, au préalable, à l'aide d'un test d'hypothèse.

13.4 Références.

On trouvera des applications de la méthode du maximum de vraisemblance en astrophysique dans, par exemple, Cash (1979) [15].

13.5 Exercices et problèmes.

Exercice 13.1. On tire au hasard un couple de nombres (X, Y) suivant la loi normale 2D de paramètres $\mu_1, \mu_2, \sigma_1, \sigma_2$ et ρ. On révèle une des deux valeurs du couple, par exemple $X = x$, donner l'estimation du maximum de vraisemblance de la valeur y prise par Y. Dans ce cas particulier la loi (marginale) de Y est connue, donner alors l'estimation correspondant au maximum de probabilité a posteriori.

Chapitre 14

La méthode des moindres carrés.

Avec l'exposé de la méthode des moindres carrés il intervient un changement de notations qu'il est bon de préciser dès maintenant. On notait jusqu'ici $(x_1, \ldots, x_n)$ les n réalisations d'un n-échantillon, mais dans l'exposé de la méthode des moindres carrés il est traditionnel de les noter $(y_1, \ldots, y_n)$, ce qui peut prêter à confusion, d'autant plus que les x_i joueront ici le rôle de constantes et non plus de variables aléatoires. De plus, dans ce chapitre, la notation $(y_1, \ldots, y_n)$ désigne indifféremment un n-échantillon ou une réalisation de cet échantillon. La confusion n'est, en général, pas possible, et le sens à donner à cette notation est précisé par le contexte.

14.1 Le principe général.

Supposons que l'on dispose d'un n-échantillon $(y_1, \ldots, y_n)$, que nous noterons comme un vecteur colonne $\boldsymbol{y}$, en adoptant la notation matricielle. Cet échantillon est issu d'une population parente à n dimensions que nous ne connaissons pas en détail (en particulier pas la loi), mais qui possède un vecteur moyenne $\mathrm{E}\{\boldsymbol{y}\}$ et une matrice des variances-covariances $\boldsymbol{V}$. Supposons également que les valeurs y_i de l'échantillon dépendent d'un ensemble de k paramètres $(\theta_1, \theta_2, \ldots, \theta_k)$, noté $\boldsymbol{\theta}$, par l'intermédiaire de n fonctions f_i, connues aux paramètres $\boldsymbol{\theta}$ près :

$$\begin{aligned} y_1 &= f_1(\boldsymbol{\theta}, \epsilon_1), \\ y_2 &= f_2(\boldsymbol{\theta}, \epsilon_2), \\ &\ldots\ldots\ldots\ldots \\ y_n &= f_n(\boldsymbol{\theta}, \epsilon_n). \end{aligned} \tag{14.1}$$

Les ϵ_i sont les composantes d'un vecteur aléatoire $\boldsymbol{\epsilon}=(\epsilon_1, \ldots, \epsilon_n)$, représentant un bruit ; c'est la présence des ϵ_i qui rend les y_i aléatoires. On admet le plus souvent que $\mathrm{E}\{\boldsymbol{\epsilon}\} = 0$, c'est-à-dire que le bruit ou les erreurs de mesure ne présentent pas de biais ou, comme on dit, d'erreur systématique.

Un cas fréquent est le cas additif qui consiste à supposer que les $f_i(\boldsymbol{\theta}, \epsilon_i)$ peuvent s'écrire sous la forme d'une fonction déterministe $\mu_i(\boldsymbol{\theta})$ plus un bruit

ϵ_i de moyenne nulle, soit :

$$y_i = f_i(\boldsymbol{\theta}, \epsilon_i) = \mu_i(\boldsymbol{\theta}) + \epsilon_i \; . \tag{14.2}$$

Les fonctions μ_i représentent un modèle connu aux k paramètres ajustables $\boldsymbol{\theta}$ près. Comme $\mathrm{E}\{\epsilon_i\} = 0$, on a $\mathrm{E}\{y_i\} = \mu_i(\boldsymbol{\theta})$. Nous appellerons indifféremment « *moyenne* » ou « *signal* » le vecteur $\boldsymbol{\mu}(\boldsymbol{\theta})$ de composantes $(\mu_1, \ldots, \mu_n)$. L'estimation de $\boldsymbol{\theta}$ revient alors à l'estimation de la moyenne de la population parente, cette moyenne étant fonction des paramètres $\boldsymbol{\theta}$.

Etant donné une réalisation $(y_1, \ldots, y_n)$ du n-échantillon constitué de n observations y_i, la méthode des moindres carrés propose de choisir l'estimation $\widehat{\boldsymbol{\theta}}$ du vecteur $\boldsymbol{\theta}$, qui minimise la dispersion des observations autour de leur valeur moyenne. On forme donc la quantité $S(\boldsymbol{\theta})$:

$$S(\boldsymbol{\theta}) = \sum_{i=1}^{n} (y_i - \mathrm{E}\{y_i\})^2 \; , \tag{14.3}$$

puis on cherche son minimum, ce qui définit implicitement la valeur de $\widehat{\boldsymbol{\theta}}$.

L'introduction de la fonction S conduit à trouver une solution acceptable « au sens des moindres carrés » du système incompatible (14.1). La fonction S est ce que l'on appelle une fonction de régularisation. La quantité $S_{\min}$ dépend du n-échantillon $(y_1, \ldots, y_n)$ et, à ce titre, est une variable aléatoire.

Le vecteur $\widehat{\boldsymbol{\theta}}$ ainsi trouvé est appelé estimateur des moindres carrés de $\boldsymbol{\theta}$. A l'estimateur $\widehat{\boldsymbol{\theta}}$ correspond un estimateur $\widehat{f_i}$ défini par $\widehat{f_i} = f_i(\widehat{\boldsymbol{\theta}}, \epsilon_i)$. Contrairement à la méthode ML, la méthode des moindres carrés n'exige pas que la densité de probabilité de la population parente soit connue. Cependant, dans le cas le plus général, les estimateurs des moindres carrés ainsi obtenus ne possèdent aucune propriété optimale, ils sont, par exemple, souvent biaisés. Il y a toutefois deux exceptions importantes que nous étudierons : le cas normal et le cas linéaire. Ces deux cas supposent un modèle additif $\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{\epsilon}$, c'est pourquoi nous allons maintenant caractériser plus finement la méthode des moindres carrés sous l'hypothèse que ce modèle est valide.

Le modèle additif $\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{\epsilon}$.

Dans le cas de ce modèle, $\mathrm{E}\{\boldsymbol{y}\} = \boldsymbol{\mu}$, et l'équation (14.3) définissant les $\widehat{\boldsymbol{\theta}}$ s'écrit alors :

$$S(\widehat{\boldsymbol{\theta}}) = \min_{\boldsymbol{\theta}} S(\boldsymbol{\theta}); \quad S(\boldsymbol{\theta}) = \sum_{i=1}^{n} (y_i - \mu_i)^2 = (\boldsymbol{y} - \boldsymbol{\mu})^t (\boldsymbol{y} - \boldsymbol{\mu}) \, . \tag{14.4}$$

Aux estimateurs $\widehat{\boldsymbol{\theta}}$ correspondent une estimation $\widehat{\boldsymbol{\mu}}$ de $\boldsymbol{\mu}$: $\widehat{\boldsymbol{\mu}} = \boldsymbol{\mu}(\widehat{\boldsymbol{\theta}})$ et une estimation $\widehat{\boldsymbol{\epsilon}}$ de $\boldsymbol{\epsilon}$ appelée *résidu*. Par définition du résidu $\widehat{\boldsymbol{\epsilon}} = \boldsymbol{y} - \widehat{\boldsymbol{\mu}}$.

C'est ce modèle $\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{\epsilon}$, qu'il conviendrait d'adopter pour résoudre par la méthode des moindres carrés le problème de la détection d'un signal $\boldsymbol{\mu}$ noyé dans un bruit de fond additif $\boldsymbol{\epsilon}$. Nous n'exigeons sur $\boldsymbol{\epsilon}$ que la connaissance de sa moyenne : $\mathrm{E}\{\boldsymbol{\epsilon}\} = 0$ et de sa matrice des variances-covariances : $\boldsymbol{V} = \mathrm{E}\{\boldsymbol{\epsilon}\boldsymbol{\epsilon}^t\} = \mathrm{E}\{(\boldsymbol{y} - \boldsymbol{\mu})(\boldsymbol{y} - \boldsymbol{\mu})^t\}$.

14.1.1 Géométrisation de la méthode des moindres carrés.

Afin de dégager des relations métriques, il est nécessaire de définir les espaces dans lesquels nos objets sont définis. Nous avons affaire à trois espaces :

1. L'espace des *observations* de dimension n que nous noterons O_n. C'est l'espace de toutes les observations possibles $\boldsymbol{y} \in O_n$.
2. L'espace des *paramètres* P_k, de dimension k. Un point $\boldsymbol{\theta}$ de cet espace définit de façon unique un modèle $\boldsymbol{\mu}(\boldsymbol{\theta})$.
3. L'espace dans lequel se répartissent les différents modèles $\boldsymbol{\mu}$. C'est un sous-espace de O_n qui est l'image par les fonctions $\boldsymbol{\mu}(\boldsymbol{\theta})$ de l'espace des paramètres P_k. Nous le noterons M_k, ($M_k = \text{ima}(\boldsymbol{\mu}(\boldsymbol{\theta})) \subseteq O_n$).

Ces espaces et sous-espaces définissent le cadre dans lequel nous devons résoudre notre problème d'estimation par la méthode des moindres carrés. Selon le principe de cette méthode, nous devons minimiser une quantité $S(\boldsymbol{\theta})$ qui est une somme de carrés dans l'espace O_n des observations. Cette remarque nous incite à munir cet espace d'une structure d'espace euclidien en introduisant un produit scalaire $(\boldsymbol{x}, \boldsymbol{y})$ entre deux éléments $\boldsymbol{x}, \boldsymbol{y}$ quelconques de O_n, que nous définissons ainsi :

$$\forall \boldsymbol{x}, \boldsymbol{y} \in O_n;\ (\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x}^t \boldsymbol{y} = \sum_{i=1}^{n} x_i y_i \,. \tag{14.5}$$

Ce produit scalaire induit une norme dite euclidienne sur les observations $\boldsymbol{y} \in O_n$, notée $||\boldsymbol{y}||$ ainsi qu'une distance d et un angle $\alpha \in [0, \pi]$ entre deux observations. On a :

$$\forall \boldsymbol{y} \in O_n;\ ||\boldsymbol{y}|| = \sqrt{(\boldsymbol{y}, \boldsymbol{y})}, \quad \text{et} \tag{14.6}$$

$$\forall \boldsymbol{x}, \boldsymbol{y} \in O_n;\ d(\boldsymbol{x}, \boldsymbol{y}) = ||\boldsymbol{x} - \boldsymbol{y}||, \cos\alpha = \frac{(\boldsymbol{x}, \boldsymbol{y})}{\sqrt{(\boldsymbol{x}, \boldsymbol{x})(\boldsymbol{y}, \boldsymbol{y})}} \,. \tag{14.7}$$

Cette norme euclidienne confère à l'espace des observations une structure d'espace euclidien. Dans cet espace, la méthode des moindres carrés devient un problème de recherche du minimum de la norme des résidus $\widehat{\boldsymbol{\epsilon}} = \boldsymbol{y} - \boldsymbol{\mu}$ ou de manière équivalente du minimum de la distance entre les observations $\boldsymbol{y}$ et une estimation $\widehat{\boldsymbol{\mu}}$ de leur moyenne :

$$S_{\min} = \min_{\boldsymbol{\theta}} S(\boldsymbol{\theta}) \iff \widehat{\boldsymbol{\epsilon}} = \min_{\boldsymbol{\theta}} ||\boldsymbol{\epsilon}(\boldsymbol{\theta})|| \iff \widehat{\boldsymbol{\mu}} = \min_{\boldsymbol{\theta}} d(\boldsymbol{y}, \boldsymbol{\mu}(\boldsymbol{\theta})) \,. \tag{14.8}$$

14.2 Le cas normal.

C'est un cas où $y_i = \mu_i(\boldsymbol{\theta}) + \epsilon_i$, et où les ϵ_i sont des variables aléatoires normales de moyenne $\mathrm{E}\{\boldsymbol{\epsilon}\} = 0$ et de matrice des variances-covariances $\mathrm{E}\{\boldsymbol{\epsilon}\boldsymbol{\epsilon}^t\} = \sigma^2 \boldsymbol{I}$, $\boldsymbol{I}$ étant la matrice identité de format (n, n). Conformément au principe de la méthode des moindres carrés, calculons la moyenne des y_i afin de former l'expression (14.3). On a $\mathrm{E}\{y_i\} = \mathrm{E}\{\mu_i(\boldsymbol{\theta}) + \epsilon_i\} = \mu_i(\boldsymbol{\theta})$ que nous noterons μ_i pour simplifier. En reportant ce résultat dans (14.4) on obtient :

$$S_{\min} = \min_{\boldsymbol{\theta}} \sum_{i=1}^{n} (y_i - \mu_i(\boldsymbol{\theta}))^2 \,. \tag{14.9}$$

Calculons maintenant le logarithme de la fonction de vraisemblance de $\boldsymbol{y}$. Pour cela, nous avons besoin de connaître la densité de probabilité des y_i. Les y_i sont des variables aléatoires normales de moyenne μ_i et de variance $\mathrm{Var}(y_i) = \mathrm{Var}(\epsilon_i) = \sigma^2$, d'où :

$$\ln L(\boldsymbol{y}|\boldsymbol{\theta}) = -\frac{1}{2} n \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu_i(\boldsymbol{\theta}))^2 \,. \tag{14.10}$$

Cette fonction est maximisée pour $\boldsymbol{\theta}$, précisément lorsque l'expression suivante est minimisée :

$$S(\boldsymbol{\theta}) = \sum_{i=1}^{n} (y_i - \mu_i(\boldsymbol{\theta}))^2 \,. \tag{14.11}$$

Les estimateurs des moindres carrés coïncident alors avec ceux du maximum de vraisemblance.

14.2.1 Moindres carrés pondérés.

A partir de la propriété précédente, on peut généraliser la méthode des moindres carrés au cas où $\boldsymbol{\epsilon}$ est un vecteur aléatoire issu d'une loi normale à n dimensions telle que :

$$\mathrm{E}\{\boldsymbol{\epsilon}\} = 0 \,, \quad \mathrm{E}\left\{\boldsymbol{\epsilon}\boldsymbol{\epsilon}^t\right\} = \sigma^2 \boldsymbol{V} \,. \tag{14.12}$$

Dans cette expression, $\sigma^2\boldsymbol{V}$ est la matrice de variances-covariances de $\boldsymbol{\epsilon}$ connue à un facteur σ^2 près, et $\boldsymbol{V}$ est la *matrice des variances-covariances relatives* de $\boldsymbol{\epsilon}$. Les y_i sont alors des variables aléatoires normales à n-dimensions, de moyenne $\mathrm{E}\{\boldsymbol{y}\} = \boldsymbol{\mu}$ et de matrice des variances-covariances $\mathrm{E}\{(\boldsymbol{y}-\boldsymbol{\mu})(\boldsymbol{y}-\boldsymbol{\mu})^t\} = \mathrm{E}\{\boldsymbol{\epsilon}\boldsymbol{\epsilon}^t\} = \sigma^2\boldsymbol{V}$. La fonction de vraisemblance s'écrit alors (voir équation (6.57)) :

$$L(\boldsymbol{y}|\boldsymbol{\theta}) = \frac{1}{(2\pi)^{\frac{n}{2}} (\sigma^2 \det \boldsymbol{V})^{\frac{1}{2}}} \exp -\frac{1}{2\sigma^2} (\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{\theta}))^t \boldsymbol{V}^{-1} (\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{\theta})) \,, \tag{14.13}$$

$$\ln L(\boldsymbol{y}|\boldsymbol{\theta}) = -\frac{1}{2} \ln[(2\pi)^n (\sigma^2 \det \boldsymbol{V})] - \frac{1}{2\sigma^2} (\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{\theta}))^t \boldsymbol{V}^{-1} (\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{\theta})) \,. \tag{14.14}$$

Le premier terme de cette expression ne dépend pas de θ et la méthode du maximum de vraisemblance conduit alors à minimiser la quantité :

$$S(\boldsymbol{\theta}) = (\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{\theta}))^t \, \boldsymbol{V}^{-1} (\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{\theta})) \,. \tag{14.15}$$

A titre d'exemple, envisageons le cas où la matrice des variances-covariances est diagonale :

$$\sigma^2\boldsymbol{V} = \sigma^2 \begin{pmatrix} \sigma_1^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_n^2 \end{pmatrix} , \quad \boldsymbol{V}^{-1} = \begin{pmatrix} \dfrac{1}{\sigma_1^2} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \dfrac{1}{\sigma_n^2} \end{pmatrix} . \tag{14.16}$$

L'expression (14.15) s'écrit à présent :

$$S(\boldsymbol{\theta}) = \sum_{i=1}^{n} \frac{(y_i - \mu_i(\boldsymbol{\theta}))^2}{\sigma_i^2} \quad . \tag{14.17}$$

Les estimateurs $\widehat{\boldsymbol{\theta}}$ du maximum de vraisemblance sont donc trouvés en minimisant l'expression :

$$S(\widehat{\boldsymbol{\theta}}) = \min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \frac{(y_i - \mu_i(\boldsymbol{\theta}))^2}{\sigma_i^2} \ , \tag{14.18}$$

qui ne fait intervenir que les variances relatives σ_i^2 des observations. La méthode qui consiste à trouver des estimateurs en minimisant une expression telle que (14.17) est appelée méthode des moindres carrés pondérés ou méthode du moindre χ^2, car la quantité $S(\widehat{\boldsymbol{\theta}})/\sigma^2$ suit la loi du χ^2 à $n-k$ degrés de liberté lorsque les y_i suivent une loi normale. Les quantités $1/\sigma_i^2$ sont souvent appelées les poids w_i des observations, car plus la précision d'une mesure y_i est grande, moins grande est sa variance, et plus grand doit être le coefficient qui rend compte de son influence dans le calcul de S. La matrice $\boldsymbol{V}^{-1}$ intervient alors comme une matrice de pondération relative qui rend compte de l'inégale précision des mesures $\boldsymbol{y}$ et de leurs corrélations éventuelles. Dans la pratique, il arrive souvent que l'on estime les poids w_i, et donc la matrice $\boldsymbol{W} = \boldsymbol{V}^{-1}$, et que l'on cherche à minimiser l'expression :

$$S(\boldsymbol{\theta}) = (\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{\theta}))^t \boldsymbol{W} (\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{\theta})) \ . \tag{14.19}$$

La solution $\widehat{\boldsymbol{\theta}}$ de cette expression est trouvée par une méthode de minimisation quelconque du genre quasi-Newton par exemple, ou comme racine du système :

$$\left(-\frac{\partial \mu(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right)^t \boldsymbol{W} (\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{\theta})) = 0 \ . \tag{14.20}$$

Mais il faut bien être conscient du fait que si $\boldsymbol{\epsilon}$ n'est pas une variable aléatoire normale ou plus généralement si sa densité de probabilité n'est pas du genre exponentiel :

$$f(x|\theta) = \exp[A(\theta)B(x) + C(x) + D(\theta)] \ , \tag{14.21}$$

il n'est absolument pas garanti d'obtenir des estimateurs optimaux. Cependant, si $\boldsymbol{W}$ est définie positive et si la matrice des variances-covariances de $\boldsymbol{y}$ reste bornée quand $n \to \infty$, alors les estimateurs des moindres carrés $\widehat{\boldsymbol{\theta}}$ sont convergents.

14.3 Le cas linéaire.

Dans ce cas, les estimateurs des moindres carrés sont optimaux même pour des échantillons de petite taille. Ils sont non-biaisés, obtenus comme combinaison linéaire des observations, et de variance minimum (MV) dans cette classe. Ces propriétés constituent le théorème de Gauss-Markov, que nous démontrerons plus loin.

14.3.1 Modèle linéaire.

Le modèle linéaire est du type $\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{\epsilon}$ où $\boldsymbol{\mu}$ s'exprime linéairement en fonction de k paramètres inconnus $\boldsymbol{\theta}$, c'est-à-dire :

$$\mu_i = \sum_{j=1}^{k} x_{ij}\theta_j, \quad i = 1, \ldots, n\,. \tag{14.22}$$

Il va sans dire que modèle linéaire ne veut pas dire ajustement par une droite des moindres carrés. L'ajustement d'un nuage de points par une parabole est un modèle linéaire. En effet l'expression : $\mu_i = \theta_0 + \theta_1 x_i + \theta_2 x_i^2$ est bien linéaire par rapport aux θ_k, comme l'exige le modèle.

Etant linéaire, le système d'équations (14.22) peut être mis sous la forme matricielle :

$$\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\theta}, \tag{14.23}$$

et le modèle linéaire lui-même, sous la forme :

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{\epsilon} \tag{14.24}$$

Rappelons la signification des termes de ces équations.

- Les vecteurs $\boldsymbol{y}$ et $\boldsymbol{\epsilon}$ sont des vecteurs colonne $(n, 1)$, représentant respectivement les données et les composantes aléatoires (le bruit). Ils appartiennent à l'espace O_n *des observations* de dimension n.

- Le vecteur $\boldsymbol{\theta}$ est un vecteur colonne $(k, 1)$, représentant les k paramètres à estimer et permettant de trouver une estimation de $\boldsymbol{\mu}$. Le vecteur $\boldsymbol{\theta}$ appartient à l'espace P_k *des paramètres*.

- Le vecteur $\boldsymbol{\mu}$ est l'image par $\boldsymbol{X}$ de $\boldsymbol{\theta}$ dans l'espace des observations O_n, c'est un vecteur colonne $(k, 1)$ qui appartient à l'espace des *modèles* M_k sous-espace de dimension k de O_n.

- La matrice $\boldsymbol{X}$ est une matrice rectangulaire (n, k) de coefficients connus appelée la « *matrice modèle* » ou encore la « *matrice de régression* ».

Structure de la matrice modèle $\boldsymbol{X}$.

Les coefficients x_{ij} du système (14.22) s'arrangent dans $\boldsymbol{X}$ sous la forme suivante :

$$\boldsymbol{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}. \tag{14.25}$$

La matrice $\boldsymbol{X}$ est formée de k colonnes de n scalaires, que nous noterons $\boldsymbol{x}_j$. L'équation $\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\theta}$ définissant le modèle peut alors s'écrire $\boldsymbol{\mu} = \sum_{j=1}^{k} \theta_j \boldsymbol{x}_j$, ce qui exprime que $\boldsymbol{\mu}$ est une combinaison linéaire des vecteurs $\boldsymbol{x}_j$. On peut alors considérer ces derniers comme une base sur laquelle on tente de décomposer $\boldsymbol{\mu}$.

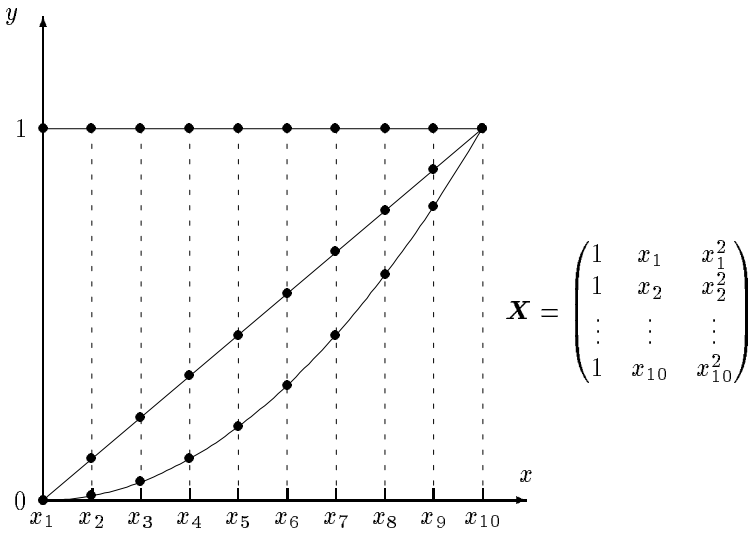


FIG. 14.1: *Construction des vecteurs colonnes de la matrice* $\boldsymbol{X}$*, dans le modèle linéaire* $y_i = \theta_1 + \theta_2 x + \theta_3 x^2$*, par échantillonnage des fonctions* $f_1(x) = 1$*,* $f_2(x) = x$ *et* $f_2(x) = x^2$*, aux points où les* y_i *sont observés.*

Les $\boldsymbol{x}_j$ ne joueront d'ailleurs ce rôle de base que s'ils sont linéairement indépendants, c'est-à-dire si l'équation $\sum_{j=1}^{k} \alpha_j \boldsymbol{x}_j = 0$ implique que les α_j sont tous nuls.

Bien souvent, on obtient les $\boldsymbol{x}_j$ par échantillonnage de fonctions continues $f_j(x)$: $x_{ij} = f_j(x_i)$. Cet échantillonnage peut être irrégulier ou régulier si $x_i = x_0 + (i-1)\Delta x$. Les observations y_i sont aussi obtenues par échantillonnage d'une certaine fonction $y(x)$ pour les mêmes x_i. Dans cette interprétation, la méthode ici exposée revient à chercher les coefficients θ_j de la combinaison linéaire des fonctions de base $\boldsymbol{x}_j$ qui approxime au mieux les données $\boldsymbol{y}$, au sens des moindres carrés.

La matrice modèle $\boldsymbol{X}$ est formée par la juxtaposition des k vecteurs colonnes $\boldsymbol{x}_j$, version échantillonnée de fonctions de base. La figure 14.1 illustre la façon d'envisager les $\boldsymbol{x}_j$ que nous venons de décrire.

14.3.2 Fonctions à estimer.

Il arrive fréquemment que l'on désire estimer d'autres paramètres β_j à partir de l'estimation des θ_i. On suppose connue la dépendance des β_j en fonction des θ_i, soit $\beta_j = \varphi_j(\boldsymbol{\theta})$, que nous noterons $\boldsymbol{\beta} = \boldsymbol{\varphi}(\boldsymbol{\theta})$. Les φ_j (et plus improprement les β_j) sont appelées les « fonctions à estimer ».

Les β_j sont en nombre s quelconque et il n'est pas nécessaire que ce nombre soit égal au nombre de paramètres θ_i. Cependant si les fonctions $\boldsymbol{\varphi}$ sont bijectives, alors on a :

$$\min_{\boldsymbol{\theta}} ||\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{\theta})|| = \min_{\boldsymbol{\beta}} ||\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{\varphi}^{-1}(\boldsymbol{\beta}))||, \tag{14.26}$$

ce qui conduit à écrire $\widehat{\boldsymbol{\beta}} = \boldsymbol{\varphi}(\widehat{\boldsymbol{\theta}})$ et l'estimation de $\widehat{\boldsymbol{\mu}}$ est indépendante de

la paramétrisation. Si les fonctions $\boldsymbol{\varphi}$ ne sont pas bijectives nous prendrons $\widehat{\boldsymbol{\beta}} = \boldsymbol{\varphi}(\widehat{\boldsymbol{\theta}})$ comme définition des estimateurs de $\boldsymbol{\beta}$ au sens des moindres carrés.

Un cas important est celui où les fonctions $\boldsymbol{\varphi}$ sont linéaires, de matrice $\boldsymbol{C}$ connue, telle que $\boldsymbol{\beta} = \boldsymbol{C\theta}$. Si $\boldsymbol{C}$ est inversible, on peut réinterpréter le cas linéaire de matrice modèle $\boldsymbol{X}$ en un autre cas linéaire de matrice modèle $\boldsymbol{X}_\beta = \boldsymbol{XC}^{-1}$. Un choix judicieux de $\boldsymbol{C}$ permet souvent de simplifier la recherche de solutions par la méthode des moindres carrés. Le théorème de Gauss-Markov établit l'optimalité des estimateurs $\widehat{\boldsymbol{\beta}}$ dans le cas linéaire, pour $\boldsymbol{\beta} = \boldsymbol{C\theta}$ et $\boldsymbol{C}$ étant (sauf cas singulier) quelconque.

14.3.3 Modèle linéaire réduit.

Le modèle linéaire réduit est un modèle linéaire pour lequel le bruit $\boldsymbol{\epsilon}$ est de moyenne nulle, de composantes non-corrélées et de variances égales, c'est-à-dire :

$$\mathrm{E}\{\boldsymbol{\epsilon}\} = 0, \qquad \mathrm{E}\left\{\boldsymbol{\epsilon\epsilon}^t\right\} = \sigma^2 \boldsymbol{I}\,. \tag{14.27}$$

On peut toujours supposer que la relation $\mathrm{E}\{\boldsymbol{\epsilon}\} = 0$ est satisfaite. Si tel n'était pas le cas, et s'il existait une *erreur systématique* μ indépendante de i ($\forall i$, $\mathrm{E}\{\epsilon_i\} = \mu$), on pourrait considérer la constante μ comme un paramètre supplémentaire du genre θ, et ré-écrire le système sous la forme :

$$\boldsymbol{y} = (\boldsymbol{X}|\mathbf{1}) \begin{pmatrix} \boldsymbol{\theta} \\ \mu \end{pmatrix} + \boldsymbol{\epsilon}'\,, \tag{14.28}$$

où $\mathrm{E}\{\boldsymbol{\epsilon}'\} = 0$ et où le $\mathbf{1}$ dans la formule précédente représente une colonne de 1 adjointe à la matrice $\boldsymbol{X}$, tandis que μ est une constante insérée sous la colonne des $\boldsymbol{\theta}$. S'il y avait plusieurs moyennes inconnues $\mu_1, \ldots, \mu_s$, il faudrait adjoindre à la matrice $\boldsymbol{X}$ autant de colonnes formées de 1 et de 0, et insérer autant de μ_i sous la colonne des $\boldsymbol{\theta}$, tout en veillant à ce qu'il n'y ait pas plus d'inconnues que d'équations, c'est-à-dire que $k + s \leq n$. Sous réserve que cette dernière condition soit remplie, il n'y a pas de restriction à étudier le système :

$$\boldsymbol{y} = \boldsymbol{X\theta} + \boldsymbol{\epsilon}\ ; \quad \mathrm{E}\{\boldsymbol{\epsilon}\} = 0\,. \tag{14.29}$$

On peut également supposer que la matrice des variances-covariances de $\boldsymbol{\epsilon}$ est de la forme :

$$\mathrm{E}\{\boldsymbol{\epsilon\epsilon}^t\} = \sigma^2 \boldsymbol{I}, \tag{14.30}$$

où la variance σ^2 est peut-être inconnue. C'est ce qui se passe lorsque les mesures $\boldsymbol{y}$ sont d'égale précision. De nouveau, si cela n'était pas le cas, et si les mesures $\boldsymbol{y}$ étaient de précision inégale et/ou corrélées, on aurait :

$$\mathrm{E}\left\{\boldsymbol{\epsilon\epsilon}^t\right\} = \sigma^2 \boldsymbol{V}\ ; \quad \boldsymbol{V} \text{ étant connue.} \tag{14.31}$$

Nous allons supposer de plus que $\boldsymbol{V}$ est non singulière, c'est-à-dire inversible ($\det \boldsymbol{V} \neq 0$). Le modèle où $\boldsymbol{V}$ est singulière a été envisagé par Kreijger et Neudecker (1977) [45]. Afin de se ramener au cas réduit, on peut alors effectuer le changement de variables $\boldsymbol{y}_0 = \boldsymbol{N}^t \boldsymbol{y}$, $\boldsymbol{X}_0 = \boldsymbol{N}^t \boldsymbol{X}$ et $\boldsymbol{\epsilon}_0 = \boldsymbol{N}^t \boldsymbol{\epsilon}$, où $\boldsymbol{N}$ est

une matrice non singulière ayant la propriété $\boldsymbol{N}\boldsymbol{N}^t = \boldsymbol{V}^{-1}$. Ce changement de variables est tel que :

$$\begin{aligned} \mathrm{E}\{\boldsymbol{\epsilon}_0\} &= \mathrm{E}\left\{\boldsymbol{N}^t\boldsymbol{\epsilon}\right\} = \boldsymbol{N}^t\,\mathrm{E}\{\boldsymbol{\epsilon}\} = 0 \\ \mathrm{E}\left\{\boldsymbol{\epsilon}_0\boldsymbol{\epsilon}_0^t\right\} &= \mathrm{E}\left\{\boldsymbol{N}^t\boldsymbol{\epsilon}\boldsymbol{\epsilon}^t\boldsymbol{N}\right\} = \boldsymbol{N}^t\,\mathrm{E}\left\{\boldsymbol{\epsilon}\boldsymbol{\epsilon}^t\right\}\boldsymbol{N} = \sigma^2\boldsymbol{N}^t\boldsymbol{V}\boldsymbol{N} \\ \boldsymbol{N}\,\mathrm{E}\left\{\boldsymbol{\epsilon}_0\boldsymbol{\epsilon}_0^t\right\} &= \sigma^2\boldsymbol{V}^{-1}\boldsymbol{V}\boldsymbol{N} = \sigma^2\boldsymbol{N} \\ \boldsymbol{N}^{-1}\boldsymbol{N}\,\mathrm{E}\left\{\boldsymbol{\epsilon}_0\boldsymbol{\epsilon}_0^t\right\} &= \mathrm{E}\left\{\boldsymbol{\epsilon}_0\boldsymbol{\epsilon}_0^t\right\} = \sigma^2\boldsymbol{I} \end{aligned}$$

Le système d'équations devient grâce à ce changement de variables :

$$\boldsymbol{y}_0 = \boldsymbol{X}_0\boldsymbol{\theta} + \boldsymbol{\epsilon}_0\ ;\ \ \mathrm{E}\{\boldsymbol{\epsilon}_0\} = 0\ \ \text{et}\ \ \mathrm{E}\left\{\boldsymbol{\epsilon}_0\boldsymbol{\epsilon}_0^t\right\} = \sigma^2\boldsymbol{I}, \tag{14.32}$$

et il est donc bien sous forme réduite. Il reste à déterminer la matrice $\boldsymbol{N}^t$. Un calcul simple montre que $\boldsymbol{N}^t = \boldsymbol{\Lambda}^{-1}\boldsymbol{U}^t$, où $\boldsymbol{\Lambda}^{-2}$ et $\boldsymbol{U}$ sont les matrices respectivement des valeurs propres et des vecteurs propres de $\boldsymbol{V}^{-1}$:

$$\boldsymbol{V}^{-1} = \boldsymbol{U}\boldsymbol{\Lambda}^{-2}\boldsymbol{U}^t\ . \tag{14.33}$$

Lorsque $n \to \infty$ on ne peut plus supposer que $\boldsymbol{V}$ reste toujours régulière et qu'un tel changement de variables est possible ; le comportement asymptotique des estimateurs des moindres carrés *pondérés* est alors subordonné au comportement asymptotique de la matrice $\boldsymbol{V}$.

14.3.4 Les équations normales.

Le changement de variables évoqué plus haut ayant été éventuellement fait, nous allons maintenant chercher la solution du modèle linéaire réduit. La méthode des moindres carrés requiert de minimiser la quantité :

$$S(\boldsymbol{\theta}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta})^t(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta})\,. \tag{14.34}$$

Si $\boldsymbol{\theta}$ est quelconque, une condition nécessaire pour que $\widehat{\boldsymbol{\theta}}$ réalise ce minimum, est qu'en $\widehat{\boldsymbol{\theta}}$ les dérivées de S par rapport aux θ_i soient nulles. Notons $\boldsymbol{S}'$ le vecteur d'éléments $\partial S/\partial\theta_i$ et $\boldsymbol{S}''$ la matrice d'éléments $\partial^2 S/\partial\theta_i\partial\theta_j$. On obtient facilement :

$$\boldsymbol{S}' = -2\boldsymbol{X}^t(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}), \qquad \boldsymbol{S}'' = 2\boldsymbol{X}^t\boldsymbol{X}\ . \tag{14.35}$$

La condition $\boldsymbol{S}' = 0$ pour $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$, impose $\boldsymbol{X}^t(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\theta}}) = 0$, ce qui conduit, afin de trouver $\widehat{\boldsymbol{\theta}}$, à résoudre le système linéaire suivant :

$$\boldsymbol{X}^t\boldsymbol{X}\widehat{\boldsymbol{\theta}} = \boldsymbol{X}^t\boldsymbol{y}\ . \tag{14.36}$$

Les équations de ce système portent le nom d'*équations normales*. Nous ne nous préoccupons pas pour le moment de savoir s'il s'agit bien d'un minimum, ce qui serait assuré si $\boldsymbol{X}^t\boldsymbol{X}$ était définie positive.

Structure de la matrice $\boldsymbol{X}^t\boldsymbol{X}$. La matrice $\boldsymbol{X}$ est formée des vecteurs de base $\boldsymbol{x}_j$ rangés sous forme de colonnes, comme discuté précédemment. La matrice $\boldsymbol{X}^t$, elle, est formée de ces mêmes vecteurs, mais écrits sous forme de lignes.

Un élément ij de $\boldsymbol{X}^t\boldsymbol{X}$ est alors le produit du vecteur ligne $\boldsymbol{x}_i^t$ par le vecteur colonne $\boldsymbol{x}_j$. On a :

$$\boldsymbol{X}^t\boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1^t\boldsymbol{x}_1 & \boldsymbol{x}_1^t\boldsymbol{x}_2 & \cdots & \boldsymbol{x}_1^t\boldsymbol{x}_k \\ \boldsymbol{x}_2^t\boldsymbol{x}_1 & \boldsymbol{x}_2^t\boldsymbol{x}_2 & \cdots & \boldsymbol{x}_2^t\boldsymbol{x}_k \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{x}_k^t\boldsymbol{x}_1 & \boldsymbol{x}_k^t\boldsymbol{x}_2 & \cdots & \boldsymbol{x}_k^t\boldsymbol{x}_k \end{pmatrix}. \tag{14.37}$$

La matrice $\boldsymbol{X}^t\boldsymbol{X}$ est donc une matrice carrée, symétrique, dont les éléments sont les produits scalaires des vecteurs de base. A la soustraction de leur moyenne près, ces produits scalaires sont les covariances des vecteurs de base, et la matrice $\boldsymbol{X}^t\boldsymbol{X}$ peut être assimilée à une matrice des variances-covariances des vecteurs de base.

En algèbre linéaire une telle matrice s'appelle la matrice de Gram de la base des vecteurs colonnes de $\boldsymbol{X}$, on pourra consulter l'ouvrage de Glazman et Liubitch [25] chap.III §3 pour connaître les diverses propriétés d'une matrice de Gram. Donnons à présent quelques propriétés des équations normales.

Les équations normales sont compatibles. Un système linéaire (comme celui associé aux équations normales $\boldsymbol{X}^t\boldsymbol{X}\widehat{\boldsymbol{\theta}} = \boldsymbol{X}^t\boldsymbol{y}$) est : soit *incompatible* et alors il n'y a aucune valeur de $\widehat{\boldsymbol{\theta}}$ susceptible d'atteindre le second membre $\boldsymbol{X}^t\boldsymbol{y}$; soit *compatible* et il y a au moins une solution $\widehat{\boldsymbol{\theta}}$ qui atteint ce second membre. Dans cette dernière circonstance, on dit que le second membre est dans l'image par $\boldsymbol{X}^t\boldsymbol{X}$ de $\widehat{\boldsymbol{\theta}}$ et on écrit $\boldsymbol{X}^t\boldsymbol{y} \in \text{ima}(\boldsymbol{X}^t\boldsymbol{X})$. Les équations normales ont ceci de particulier qu'elles sont obligatoirement *compatibles*.

Pour qu'un système linéaire soit compatible, il faut et il suffit qu'il satisfasse la condition de Fredholm qui exige que toute solution $\boldsymbol{z}$ du système homogène adjoint : $(\boldsymbol{X}^t\boldsymbol{X})^t\boldsymbol{z} = \boldsymbol{0}$, soit orthogonale au second membre, c'est-à-dire $\boldsymbol{z}^t\boldsymbol{X}^t\boldsymbol{y} = 0$. Dans le cas des équations normales, la matrice $\boldsymbol{X}^t\boldsymbol{X}$ est symétrique et $\boldsymbol{z}$ est aussi solution du système homogène $\boldsymbol{X}^t\boldsymbol{X}\boldsymbol{z} = \boldsymbol{0}$, ce qui exprime que $\boldsymbol{z}$ appartient au noyau de $\boldsymbol{X}^t\boldsymbol{X}$ ($\boldsymbol{z} \in \ker(\boldsymbol{X}^t\boldsymbol{X})$). Si $\boldsymbol{z}$ est réduit au seul élément nul ($\ker(\boldsymbol{X}^t\boldsymbol{X}) = \{0\}$) la condition de Fredholm est automatiquement satisfaite, si le noyau n'est pas réduit au seul élément nul il vient :

$$\begin{aligned} \boldsymbol{X}^t\boldsymbol{X}\boldsymbol{z} = \boldsymbol{0} \;\Rightarrow\; \boldsymbol{z}^t\boldsymbol{X}^t\boldsymbol{X}\boldsymbol{z} = \boldsymbol{0} \;\Rightarrow\; ||\boldsymbol{X}\boldsymbol{z}|| = 0 \;\Rightarrow\; \boldsymbol{X}\boldsymbol{z} = \boldsymbol{0} \;\Rightarrow\; \boldsymbol{X}^t\boldsymbol{X}\boldsymbol{z} = \boldsymbol{0} \\ \boldsymbol{X}\boldsymbol{z} = \boldsymbol{0} \;\Rightarrow\; (\boldsymbol{X}\boldsymbol{z})^t\boldsymbol{y} = 0 \;\Rightarrow\; \boldsymbol{z}^t\boldsymbol{X}^t\boldsymbol{y} = 0\,. \end{aligned} \tag{14.38}$$

La dernière égalité est la condition de Fredholm et les équations normales sont bien compatibles. Nous avons comme conséquence directe de (14.38) :

Corollaire 14.1. *Le noyau de $\boldsymbol{X}^t\boldsymbol{X}$ est confondu avec le noyau de $\boldsymbol{X}$, c'est-à-dire :*

$$\boldsymbol{X}^t\boldsymbol{X}\boldsymbol{z} = \boldsymbol{0} \iff \boldsymbol{X}\boldsymbol{z} = \boldsymbol{0} \tag{14.39}$$

Un élément $\boldsymbol{z}$ du noyau de $\boldsymbol{X}$ exprime la dépendance linéaire des colonnes de $\boldsymbol{X}$, c'est-à-dire des fonctions de bases échantillonnées.

14.3.5 Solution du modèle linéaire.

Nous sommes maintenant en mesure d'énoncer le théorème concernant l'estimation de $\boldsymbol{\theta}$ par $\widehat{\boldsymbol{\theta}}$ au sens des moindres carrés.

Théorème 14.2. *Si $\boldsymbol{\theta}$ est un vecteur de paramètres susceptibles de prendre des valeurs quelconques, alors le vecteur $\widehat{\boldsymbol{\theta}}$ est un estimateur de $\boldsymbol{\theta}$ au sens des moindres carrés si, et seulement si, $\widehat{\boldsymbol{\theta}}$ est solution des équations normales. C'est-à-dire :*

$$\forall \widetilde{\boldsymbol{\theta}},\ \min_{\widetilde{\boldsymbol{\theta}}} S(\widetilde{\boldsymbol{\theta}}) = S(\widehat{\boldsymbol{\theta}}) \iff \boldsymbol{X}^t\boldsymbol{X}\widehat{\boldsymbol{\theta}} = \boldsymbol{X}^t\boldsymbol{y}, \qquad (14.40)$$
$$\text{où} \quad S(\widetilde{\boldsymbol{\theta}}) = (\boldsymbol{y} - \boldsymbol{X}\widetilde{\boldsymbol{\theta}})^t(\boldsymbol{y} - \boldsymbol{X}\widetilde{\boldsymbol{\theta}})\,.$$

Démonstration. Nous avons déjà montré que si $\widehat{\boldsymbol{\theta}}$ était un estimateur de $\boldsymbol{\theta}$ au sens des moindres carrés alors il était nécessairement solution des équations normales. Il reste à démontrer que si $\widehat{\boldsymbol{\theta}}$ est solution des équations normales, alors c'est aussi une solution de $\min_{\widetilde{\theta}} S(\widetilde{\boldsymbol{\theta}})$. Considérons un autre estimateur quelconque $\widehat{\boldsymbol{\theta}} + \Delta\boldsymbol{\theta}$. Il vient :

$$\begin{aligned} S(\widehat{\boldsymbol{\theta}} + \Delta\boldsymbol{\theta}) &= (\boldsymbol{y} - \boldsymbol{X}(\widehat{\boldsymbol{\theta}} + \Delta\boldsymbol{\theta}))^t(\boldsymbol{y} - \boldsymbol{X}(\widehat{\boldsymbol{\theta}} + \Delta\boldsymbol{\theta})) = \\ &(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\theta}})^t(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\theta}}) - 2(\Delta\boldsymbol{\theta})^t\boldsymbol{X}^t(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\theta}}) + (\Delta\boldsymbol{\theta})^t\boldsymbol{X}^t\boldsymbol{X}\Delta\boldsymbol{\theta} = \\ &S(\widehat{\boldsymbol{\theta}}) - 2(\Delta\boldsymbol{\theta})^t(\boldsymbol{X}^t\boldsymbol{y} - \boldsymbol{X}^t\boldsymbol{X}\widehat{\boldsymbol{\theta}}) + (\Delta\boldsymbol{\theta})^t\boldsymbol{X}^t\boldsymbol{X}\Delta\boldsymbol{\theta}\,. \end{aligned}$$

Le deuxième terme du dernier membre est nul car $\widehat{\boldsymbol{\theta}}$ est solution des équations normales ; le troisième terme n'est pas négatif car c'est un carré. Il vient alors $S(\widehat{\boldsymbol{\theta}}+\Delta\boldsymbol{\theta}) \geq S(\widehat{\boldsymbol{\theta}})$ et $\widehat{\boldsymbol{\theta}}$ est bien solution de $\min_{\widetilde{\theta}} S(\widetilde{\boldsymbol{\theta}})$. □

La condition « $\boldsymbol{\theta}$ *quelconque* » n'est pas anodine : si $\boldsymbol{\theta}$ est restreint à un certain domaine de l'espace des paramètres, il est possible que la solution des équations normales soit en dehors de ce domaine et on n'obtient alors pas l'estimateur des moindres carrés comme solution des équations normales.

On sait que la solution générale d'un système linéaire est égale à une solution particulière de l'équation *avec* second membre, plus la solution générale du système *sans* second membre (système homogène). Toutes ces solutions forment un espace vectoriel $\mathcal{N}$ dont la dimension est égale à la dimension du noyau. C'est le théorème bien connu en algèbre linéaire sous le nom de théorème de la « translation du noyau ». On peut alors écrire toutes les solutions des équations normales sous la forme :

$$\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}_0 + \boldsymbol{z} \quad \text{avec} \quad \boldsymbol{X}^t\boldsymbol{X}\widehat{\boldsymbol{\theta}}_0 = \boldsymbol{X}^t\boldsymbol{y},\ \text{et}\ \boldsymbol{X}^t\boldsymbol{X}\boldsymbol{z} = 0\,. \qquad (14.41)$$

Il n'y aura de solution unique que si la solution de $\boldsymbol{X}^t\boldsymbol{X}\boldsymbol{z} = 0$ est réduite à la seule solution triviale $\boldsymbol{z} = 0$. On a vu en (14.38) que la condition $\boldsymbol{X}^t\boldsymbol{X}\boldsymbol{z} = 0$ est équivalente à $\boldsymbol{X}\boldsymbol{z} = 0$, cette équation n'admet pour solution que la solution triviale que si les colonnes de $\boldsymbol{X}$, qui sont les fonctions de bases, sont linéairement indépendantes. On a alors :

Théorème 14.3. *Les équations normales possèdent une solution unique si, et seulement si, le système $\boldsymbol{X}\boldsymbol{z} = 0$ n'admet que la seule solution : $\boldsymbol{z} = 0$, c'est-à-dire si les fonctions de base qui forment les colonnes de la matrice modèle $\boldsymbol{X}$ sont linéairement indépendantes.*

Cas où la solution est unique (cas régulier).

Si les équations normales admettent une solution unique, l'estimateur de $\boldsymbol{\theta}$ au sens des moindres carrés : $\widehat{\boldsymbol{\theta}}$ est donné par l'expression :

$$\widehat{\boldsymbol{\theta}} = (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{y}\,. \tag{14.42}$$

La matrice $\boldsymbol{X}^t\boldsymbol{X}$ est définie positive et avec elle $\boldsymbol{S''}$. On retrouve ainsi que la solution (14.42) correspond bien à un minimum. Dans le cas non réduit où $\boldsymbol{X}$ doit être remplacé par $\boldsymbol{N}^t\boldsymbol{X}$ et $\boldsymbol{y}$ par $\boldsymbol{N}^t\boldsymbol{y}$ il vient, en se rappelant que $\boldsymbol{N}\boldsymbol{N}^t = \boldsymbol{V}^{-1}$:

$$\boldsymbol{S''} = 2\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X} \tag{14.43}$$

$$\widehat{\boldsymbol{\theta}} = inv(\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X})\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{y} \tag{14.44}$$

ou encore, en posant $\boldsymbol{V}^{-1} = \boldsymbol{W}$ (la matrice des poids relatifs), on obtient :

$$\boldsymbol{S''} = 2\boldsymbol{X}^t\boldsymbol{W}\boldsymbol{X} \tag{14.45}$$

$$\widehat{\boldsymbol{\theta}} = (\boldsymbol{X}^t\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{W}\boldsymbol{y}\,. \tag{14.46}$$

L'estimateur $\widehat{\boldsymbol{\theta}}$ est une combinaison linéaire des observations $\boldsymbol{y}$: c'est un estimateur *linéaire* et la méthode des moindres carrés est un cas particulier de l'estimation linéaire. Cette théorie traite de l'estimation de k valeurs à partir d'une combinaison linéaire de n observations. Dans le cas qui nous intéresse ici on a $k < n$. C'est ce qu'on appelle une régression ou un lissage.

▶**Exemple 14.1.** *Estimation d'une constante.* Prenons le cas le plus simple de l'estimation d'un seul paramètre θ, à partir d'un ensemble de n mesures y_i. Le modèle correspondant est donc : $\forall i \quad i = 1 \ldots n,\ y_i = \theta + \epsilon_i$, que l'on notera : $\boldsymbol{y} = \mathbf{1}\theta + \boldsymbol{\epsilon}$. Le symbole $\mathbf{1}$ représente un vecteur colonne $(n, 1)$ formé uniquement de 1. On a donc $\boldsymbol{X} = \mathbf{1}$, et :

$$\boldsymbol{X}^t\boldsymbol{X} = \mathbf{1}^t\mathbf{1} = n, \quad (\boldsymbol{X}^t\boldsymbol{X})^{-1} = \frac{1}{n}, \quad \boldsymbol{X}^t\boldsymbol{y} = \mathbf{1}^t\boldsymbol{y} = \sum_{i=1}^{n} y_i\,,$$

$$\text{d'où l'estimation de } \theta : \quad \widehat{\theta} = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y}\,. \tag{14.47}$$

L'estimateur des moindres carrés $\widehat{\boldsymbol{\theta}}$: n'est autre que la moyenne arithmétique des y_i.

▶**Exemple 14.2.** *Mesures d'inégale précision.* Supposons que les mesures soient d'inégale précision, affectées d'un poids relatif $w_i = 1/\sigma_i^2$ si les σ_i^2 sont connus, ou de tout autre poids w_i jugé raisonnable par l'expérimentateur. En conservant le modèle $\boldsymbol{y} = \mathbf{1}\theta + \boldsymbol{\epsilon}$ de l'exemple précédent, on aura :

$$(\mathbf{1}^t\boldsymbol{W}\mathbf{1})^{-1} = \Big(\sum_{i=1}^{n} w_i\Big)^{-1}, \quad \mathbf{1}^t\boldsymbol{W}\boldsymbol{y} = \sum_{i=1}^{n} w_i y_i$$

$$\text{d'où} \quad \widehat{\theta} = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i}\,. \tag{14.48}$$

La valeur $\widehat{\theta}$ représente le centre de gravité des y_i affectés des poids w_i.

Bases orthonormées. Le calcul des solutions (14.42) serait grandement simplifié si les vecteurs colonnes de $\boldsymbol{X}$ étaient orthonormés. Dans cette éventualité, on aurait $\boldsymbol{X}^t\boldsymbol{X} = \boldsymbol{I}_k$ et il n'y aurait pas d'inverse à calculer : $\boldsymbol{\theta}$ serait égal à $\boldsymbol{X}^t\boldsymbol{y}$ et $\widehat{\boldsymbol{\mu}}$ à $\boldsymbol{X}\boldsymbol{X}^t\boldsymbol{y}$. On peut atteindre cet objectif avec une autre paramétrisation de $\boldsymbol{\mu}$ où les nouveaux paramètres $\boldsymbol{\beta}$ sont tels que la matrice du modèle correspondant $\boldsymbol{X}_\beta$ soit orthonormée. Si l'on a pas connaissance *a priori* d'une base orthonormée, il est toujours possible d'en obtenir une à partir d'une base quelconque à l'aide, par exemple, du processus d'orthogonalisation de Gram-Schmidt (voir annexe B.3.5, page 323).

Ce travail étant fait, et si c'est seulement l'estimation $\widehat{\boldsymbol{\mu}}$ du signal $\boldsymbol{\mu}$ qui importe (et non pas celle des paramètres), on l'obtiendra par $\widehat{\boldsymbol{\mu}} = \boldsymbol{X}_\beta\widehat{\boldsymbol{\beta}}$; $\widehat{\boldsymbol{\beta}} = \boldsymbol{X}_\beta^t\boldsymbol{y}$. Si néanmoins on désire aussi une estimation des $\boldsymbol{\theta}$ on l'obtiendra par $\widehat{\boldsymbol{\theta}} = \boldsymbol{A}\widehat{\boldsymbol{\beta}}$.

▶**Exemple 14.3.** *Base orthogonale pour un échantillonnage régulier.* Si X désigne un ensemble de $2n$ points espacés *régulièrement*:

$$X = \{x_i \,|\, x_i = \frac{i\pi}{n},\ i = -n+1, \ldots, -1, 0, 1, 2, \ldots, n\},$$

alors les fonctions trigonométriques échantillonnées sur X sont orthogonales :

$$\sum_{x\in X} \sin jx \sin kx = 0\,, \quad j \neq k;$$
$$\sum_{x\in X} \sin jx \cos kx = 0\,;$$
$$\sum_{x\in X} \cos jx \cos kx = 0\,, \quad j \neq k.$$

Cas où il y a plusieurs solutions (cas singulier).

Avant d'exposer la méthode qui permet de traiter un problème singulier de ce type, il importe de comprendre pourquoi on peut se retrouver dans une telle situation. D'après le théorème 14.3 il y a plusieurs solutions si, et seulement si, les colonnes de $\boldsymbol{X}$ ne sont pas linéairement indépendantes, ce qui peut arriver à la suite des circonstances suivantes.

- Il y a plus de fonctions de base que de points expérimentaux : les colonnes de $\boldsymbol{X}$ sont alors nécessairement linéairement dépendantes.

- Les fonctions de bases n'ont pas été choisies avec soin et il se trouve qu'au moins l'une d'elles est linéairement dépendante des autres. Supposons, par exemple, que l'on ait choisi de représenter un signal périodique $y(t)$ de période T à l'aide du modèle $\mu(t)$:

$$\mu(t) = \theta_0 + \theta_1 \cos\Big(\frac{2\pi}{T}t\Big) + \theta_2 \cos\Big(\frac{4\pi}{T}t\Big) + \theta_3 \cos^2\Big(\frac{2\pi}{T}t\Big)\,.$$

  Les fonctions de base sont : 1, $\cos(2\pi t/T)$, $\cos(4\pi t/T)$ et $\cos^2(2\pi t/T)$. Puisque $2\cos^2(2\pi t/T) = 1 + \cos(2\pi t/T)$, on est assuré, avant même tout échantillonnage du signal et des fonctions de base qui nous délivreraient respectivement le vecteur $\boldsymbol{y}$ et la matrice modèle $\boldsymbol{X}$, que la quatrième colonne de $\boldsymbol{X}$ sera linéairement dépendante de la première et de la troisième.

- Les fonctions de base sont linéairement indépendantes, mais l'échantillonnage n'est pas assez serré ou est incapable de les distinguer comme telles. Les fonctions de base deviennent linéairement dépendantes vis-à-vis de l'échantillonnage. Si, dans l'exemple précédent, on choisit de représenter le signal périodique $y(t)$ par le modèle suivant :

$$\mu(t) = \theta_0 + \theta_1 \cos\Big(\frac{2N\pi}{T}t\Big) + \theta_2 \sin\Big(\frac{2N\pi}{T}t\Big)\,,$$

  tout échantillonnage *régulier* dont le pas Δt sera un multiple de $T/2N$ donnera des valeurs nulles pour le sinus et égales à un ou moins un pour le cosinus, et la matrice $\boldsymbol{X}$ sera singulière. Le pas d'échantillonnage critique $\Delta t_c = T/2N$ pour N fixé s'appelle le *pas de Shannon*. La fréquence d'échantillonnage $1/\Delta t$ doit être supérieure à deux fois la fréquence maximale N/T présente dans le modèle [1].

- Enfin il est possible que les fonctions de base soient en théorie linéairement indépendantes vis-à-vis de l'échantillonnage, mais qu'en pratique (numériquement) elles ne le soient pas. Ce cas est très répandu lorsque le nombre de fonctions de base et par conséquent le format de la matrice $\boldsymbol{X}^t\boldsymbol{X}$ deviennent très grands et que le rapport entre la plus petite et la plus grande valeur propre de $\boldsymbol{X}^t\boldsymbol{X}$ dépasse les capacités de la machine.

Il reste maintenant à caractériser les solutions des équations normales qui, rappelons-le, ne sont connues qu'à une solution du système homogène près. La solution générale du système homogène n'est autre qu'une combinaison linéaire quelconque d'une base du noyau. Il existe plusieurs solutions techniques pour obtenir une base du noyau, l'une des plus courantes est d'avoir recours à la « décomposition en valeurs singulières » de $\boldsymbol{X}^t\boldsymbol{X}$ (voir par exemple Press *et al.* [60] chap.2.9).

Pour la solution particulière de l'équation avec second membre, on choisit parmi toutes les solutions possibles la solution $\widehat{\boldsymbol{\theta}}_1$ dont la norme euclidienne est la plus petite. Cette solution s'appelle la *pseudo-solution normale* du système (en général incompatible) $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta}$. Par définition de la norme euclidienne, $\widehat{\boldsymbol{\theta}}_1$ est solution de :

$$||\widehat{\boldsymbol{\theta}}_1|| = \widehat{\boldsymbol{\theta}}_1^{\,t}\widehat{\boldsymbol{\theta}}_1 = \min_{\widehat{\boldsymbol{\theta}}} \widehat{\boldsymbol{\theta}}^{\,t}\widehat{\boldsymbol{\theta}}; \quad \text{avec} \quad \boldsymbol{X}^t\boldsymbol{X}\widehat{\boldsymbol{\theta}} = \boldsymbol{X}^t\boldsymbol{y}\,. \tag{14.49}$$

Le théorème suivant caractérise la pseudo-solution normale.

Théorème 14.4. *La pseudo-solution normale d'un système linéaire $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta}$ est l'unique solution des équations normales associées $\boldsymbol{X}^t\boldsymbol{X}\boldsymbol{\theta} = \boldsymbol{X}^t\boldsymbol{y}$ qui possède l'une quelconque des propriétés suivantes :*

0. C'est par définition la solution des équations normales qui possède la plus petite norme euclidienne.

1. C'est l'unique solution des équations normales qui soit de la forme $\boldsymbol{X}^t\boldsymbol{b}$ où $\boldsymbol{b}$ est un vecteur quelconque de l'espace O_n des observations.

1. On montre (théorème de Shannon) que pour garder *toute* l'information contenue dans les données, il faut échantillonner régulièrement à une fréquence supérieure à deux fois la fréquence maximum contenue dans les *données* et non pas seulement dans le modèle (Whittaker 1945 [70], Shannon 1949 [66]).

2. *Elle est la projection orthogonale des solutions des équations normales sur l'espace « semblable au second membre» des vecteurs de la forme* $\boldsymbol{X}^t\boldsymbol{b}$.

Démonstration. Soit $\mathcal{N}$ l'espace des solutions des équations normales. Puisque ces équations sont compatibles on a d'après (14.38) $\boldsymbol{z}^t\boldsymbol{X}^t\boldsymbol{y} = 0$, ce qui exprime que le vecteur $\boldsymbol{X}^t\boldsymbol{y}$ et (comme $\boldsymbol{y}$ est quelconque) tous les vecteurs de la forme $\boldsymbol{X}^t\boldsymbol{b}$ sont orthogonaux au noyau de $\boldsymbol{X}^t\boldsymbol{X}$. Soit $\mathcal{I}$ l'espace des vecteurs de la forme $\boldsymbol{X}^t\boldsymbol{b}$ où $\boldsymbol{b}$ est un vecteur quelconque de O_n. La compatibilité des équations normales s'exprime par $\mathcal{I} \perp \ker(\boldsymbol{X}^t\boldsymbol{X})$.

D'après les résultats élémentaires d'algèbre linéaire, on sait que tout vecteur est la somme unique d'un vecteur d'un certain sous-espace et d'un vecteur du sous-espace qui lui est orthogonal. Appliquons ce résultat aux solutions des équations normales : $\widehat{\boldsymbol{\theta}} \in \mathcal{N}$ se décompose de façon *unique* en $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}_1 + \boldsymbol{z}$ où $\widehat{\boldsymbol{\theta}}_1 \in \mathcal{I}$ et $\boldsymbol{z} \in \ker(\boldsymbol{X}^t\boldsymbol{X})$. Pour une autre solution $\widehat{\boldsymbol{\theta}}' \in \mathcal{N}$, on a $\widehat{\boldsymbol{\theta}}' = (\widehat{\boldsymbol{\theta}}' - \widehat{\boldsymbol{\theta}} + \boldsymbol{z}) + \widehat{\boldsymbol{\theta}}_1$. Mais $\widehat{\boldsymbol{\theta}}' - \widehat{\boldsymbol{\theta}} + \boldsymbol{z} \in \ker \boldsymbol{X}^t\boldsymbol{X}$ car $\widehat{\boldsymbol{\theta}}'$ et $\widehat{\boldsymbol{\theta}}$ sont des solutions des équations normales, $\boldsymbol{z} \in \ker \boldsymbol{X}^t\boldsymbol{X}$ et par conséquent $\widehat{\boldsymbol{\theta}}_1$ est la projection d'une solution quelconque des équations normales sur l'espace $\mathcal{I}$ des vecteurs de la forme $\boldsymbol{X}^t\boldsymbol{b}$.

Nous voulons à présent montrer que $||\widehat{\boldsymbol{\theta}}_1|| \leq ||\widehat{\boldsymbol{\theta}}||$. On a $||\widehat{\boldsymbol{\theta}}|| = ||\widehat{\boldsymbol{\theta}}_1 + \boldsymbol{z}|| = ||\widehat{\boldsymbol{\theta}}_1|| + ||z||$ car $\widehat{\boldsymbol{\theta}}_1 \perp \boldsymbol{z}$. Par définition une norme est positive ou nulle, en particulier $||\boldsymbol{z}|| \geq 0$ et donc $||\widehat{\boldsymbol{\theta}}_1|| \leq ||\widehat{\boldsymbol{\theta}}||$, l'égalité n'ayant lieu que si $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}_1$. □

▶**Exemple 14.4.** Soit le système d'une équation à deux inconnues $\theta_1 + \theta_2 = 2$. La matrice du modèle est $\boldsymbol{X} = (1, 1)$ et les équations normales associées sont :

$$\begin{pmatrix}1\\1\end{pmatrix}\begin{pmatrix}1 & 1\end{pmatrix}\begin{pmatrix}\theta_1\\\theta_2\end{pmatrix} = \begin{pmatrix}1\\1\end{pmatrix} 2\,. \tag{14.50}$$

La solution générale de ce système est

$$\begin{pmatrix}\theta_1\\\theta_2\end{pmatrix} = \begin{pmatrix}2\\0\end{pmatrix} + \alpha\begin{pmatrix}-1\\+1\end{pmatrix}, \tag{14.51}$$

où α est un nombre quelconque. La pseudo-solution normale est d'après 14.4 la seule solution de (14.51) qui soit semblable au second membre de (14.50). Cette solution est obtenue pour $\alpha = 1$ et la pseudo-solution normale de $\theta_1 + \theta_2 = 2$ est alors $\theta_1 = 1$ et $\theta_2 = 1$.

Matrice pseudo-inverse. Il est commode d'exprimer les solutions d'un système singulier en termes de matrice pseudo-inverse. Une matrice pseudo-inverse ou inverse de Moore-Penrose d'une matrice $\boldsymbol{A}$ est l'inverse de $\boldsymbol{A}$ dans le sous-espace où elle est inversible. Pratiquement, si l'on veut une solution du système $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ où $\boldsymbol{A}$ est singulière de rang $r \neq 0$, on projette orthogonalement $\boldsymbol{b}$ sur le sous-espace de dimension r des vecteurs propres de $\boldsymbol{A}$, on résout le système dans ce sous-espace, et on étend la solution à l'espace originel en complètant avec des zéros. (voir par exemple Glazman et Liubitch [25] chap.IX §3). Pour notre part nous adopterons la définition suivante.

Définition 14.1. La matrice pseudo-inverse d'une matrice $\boldsymbol{A}$ de format (n, k) est une matrice notée $\boldsymbol{A}^{(-1)}$ de format (k, n), dont les colonnes $\boldsymbol{x}_i$ sont les pseudo-solutions normales des k équations linéaires à n inconnues de la forme :

$$\boldsymbol{A}\boldsymbol{x}_i = \boldsymbol{e}_i, \quad i = 1, \ldots, k\,, \tag{14.52}$$

où les $\boldsymbol{e}_i$ sont les colonnes de la matrice unité d'ordre k. Les pseudo-solutions normales étant uniques, la matrice pseudo-inverse est également unique.

Propriétés. Nous donnons sans démonstration quelques propriétés des matrices pseudo-inverses, on se reportera par exemple à Beklémichev [5] chap.14 ou à l'ouvrage d'Albert [3]. Pour les propriétés des projecteurs on consultera Glazman & Liubitch [25] chap.II §8.

1. La matrice pseudo-inverse $\boldsymbol{A}^{(-1)}$ est une inverse généralisée, c'est-à-dire qu'elle possède la propriété :

$$\boldsymbol{A}\boldsymbol{A}^{(-1)}\boldsymbol{A} = \boldsymbol{A}. \tag{14.53}$$

2. La matrice pseudo-inverse est la seule inverse généralisée qui possède aussi les propriétés suivantes :

$$\boldsymbol{A}^{(-1)}\boldsymbol{A}\boldsymbol{A}^{(-1)} = \boldsymbol{A}^{(-1)}, \tag{14.54}$$

$$\left(\boldsymbol{A}\boldsymbol{A}^{(-1)}\right)^t = \boldsymbol{A}\boldsymbol{A}^{(-1)}, \quad \left(\boldsymbol{A}^{(-1)}\boldsymbol{A}\right)^t = \boldsymbol{A}^{(-1)}\boldsymbol{A}\,. \tag{14.55}$$

3. La matrice $\boldsymbol{Q} = \boldsymbol{A}^{(-1)}\boldsymbol{A}$ est un projecteur, c'est-à-dire $\boldsymbol{Q}^2 = \boldsymbol{Q}$. D'après (14.55), c'est aussi une matrice symétrique.

4. La matrice $\boldsymbol{P}$ *supplémentaire* de $\boldsymbol{Q} = \boldsymbol{A}^{(-1)}\boldsymbol{A}$, définie par :

$$\boldsymbol{P} = \boldsymbol{I}_k - \boldsymbol{A}^{(-1)}\boldsymbol{A}, \tag{14.56}$$

où $\boldsymbol{I}_k$ est l'identité dans l'espace des paramètres P_k, est une matrice symétrique et est un projecteur sur le noyau de $\boldsymbol{A}$. Les projecteurs $\boldsymbol{P}$ et $\boldsymbol{Q}$ sont, (comme tout projecteurs supplémentaires), orthogonaux entre eux, c'est-à-dire que $\boldsymbol{P}\boldsymbol{Q} = \boldsymbol{Q}\boldsymbol{P} = \boldsymbol{0}$.

5. On a la relation duale :

$$\left(\boldsymbol{A}^{(-1)}\right)^t = \left(\boldsymbol{A}^t\right)^{(-1)}, \tag{14.57}$$

avec des propriétés équivalentes aux relations 3 et 4 pour les projecteurs associés.

Les propriétés qui vont suivre se rapportent à la résolution d'un système linéaire $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ de n équations à k inconnues. A ce système est associé le système normal $\boldsymbol{A}^t\boldsymbol{A}\boldsymbol{x} = \boldsymbol{A}^t\boldsymbol{b}$ qui, on le sait, est compatible. On note $\mathcal{K}$ le noyau de $\boldsymbol{A}$, c'est-à-dire l'ensemble des solutions du système homogène $\boldsymbol{A}\boldsymbol{z} = \boldsymbol{0}$, cet espace est identique au noyau de $\boldsymbol{A}^t\boldsymbol{A}$. On note $\mathcal{I}$ l'ensemble des vecteurs semblables au second membre des équations normales, c'est-à-dire de la forme $\boldsymbol{A}^t\widetilde{\boldsymbol{b}}$, où $\widetilde{\boldsymbol{b}}$ est quelconque. Nous supposons de plus que tous les espaces de vecteurs sont munis du produit scalaire $(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x}^t\boldsymbol{y}$.

La propriété de compatibilité des équations normales nous dit que les espaces $\mathcal{I}$ et $\mathcal{K}$ sont orthogonaux (pour le produit scalaire défini plus haut) et que par conséquent un vecteur $\boldsymbol{x}$ quelconque se décompose de façon *unique* en une somme $\boldsymbol{x} = \boldsymbol{x}_1 + \boldsymbol{x}_0$ où $\boldsymbol{x}_1 \in \mathcal{I}$ et $\boldsymbol{x}_0 \in \mathcal{K}$ avec $(\boldsymbol{x}_1, \boldsymbol{x}_0) = 0$.

6. La matrice $\boldsymbol{Q}$ est un projecteur *orthogonal* de $\boldsymbol{x}$ sur l'espace $\mathcal{I}$. Nous savions déja que c'est un projecteur il est de plus orthogonal sur $\mathcal{I}$ c'est--à-dire que si $\boldsymbol{x} \in \mathcal{I}$ et $\boldsymbol{z}$ est tel que $\boldsymbol{Q}\boldsymbol{z} = 0$, alors $(\boldsymbol{z}, \boldsymbol{x}) = 0$. Cela est conforme à la définition d'un projecteur orthogonal qui exige que son noyau soit orthogonal à son image.

7. Le projecteur $\boldsymbol{P}$ projette également $\boldsymbol{x}$ orthogonalement sur $\mathcal{K}$.

8. Le vecteur $\boldsymbol{x}_1 = \boldsymbol{A}^{(-1)}\boldsymbol{b}$ est la pseudo-solution normale du système $\boldsymbol{Ax} = \boldsymbol{b}$. Si ce système est compatible sa solution générale est alors :

$$\boldsymbol{x} = \boldsymbol{A}^{(-1)}\boldsymbol{b} + (\boldsymbol{I}_k - \boldsymbol{A}^{(-1)}\boldsymbol{A})\boldsymbol{h}, \quad \boldsymbol{h} \in \mathbb{R}^k. \tag{14.58}$$

9. Si le système est incompatible ses solutions au sens des moindres carrés sont données par (14.58), mais on a aussi :

$$\boldsymbol{x} = (\boldsymbol{A}^t\boldsymbol{A})^{(-1)}\boldsymbol{b} + (\boldsymbol{I}_k - \boldsymbol{A}^{(-1)}\boldsymbol{A})\boldsymbol{h}, \quad \boldsymbol{h} \in \mathbb{R}^k. \tag{14.59}$$

Construction. Nous donnons maintenant, toujours sans démonstration, différentes méthodes permettant de calculer une matrice pseudo-inverse. On consultera Beklémichev [5] chap.XIV §3, pour obtenir les démonstrations des propriétés que nous allons donner.

1. Si les colonnes de la matrice $\boldsymbol{A}$ sont linéairement indépendantes, alors :

$$\boldsymbol{A}^{(-1)} = (\boldsymbol{A}^t\boldsymbol{A})^{-1}\boldsymbol{A}^t \tag{14.60}$$

2. Si les lignes de la matrice $\boldsymbol{A}$ sont linéairement indépendantes, alors :

$$\boldsymbol{A}^{(-1)} = \boldsymbol{A}^t(\boldsymbol{A}\boldsymbol{A}^t)^{-1} \tag{14.61}$$

3. Il est toujours possible de décomposer une matrice $\boldsymbol{A}$ de format (n, k) et de rang $r \leq \min(k, n)$ en un produit de deux matrices de rang r $\boldsymbol{BC}$, où $\boldsymbol{B}$ est de format (n, r) et $\boldsymbol{C}$ de format (r, k). Une telle décomposition est appelée *décomposition squelettique*. Si $\boldsymbol{A} = \boldsymbol{BC}$ est une décomposition squelettique de $\boldsymbol{A}$, alors :

$$\boldsymbol{A}^{(-1)} = \boldsymbol{C}^{(-1)}\boldsymbol{B}^{(-1)} = \boldsymbol{C}^t(\boldsymbol{C}\boldsymbol{C}^t)^{-1}(\boldsymbol{B}^t\boldsymbol{B})^{-1}\boldsymbol{B}^t. \tag{14.62}$$

4. Il est aussi possible de décomposer la matrice $\boldsymbol{A}$ en un produit de trois matrices $\boldsymbol{A} = \boldsymbol{UDV}^t$, où $\boldsymbol{D}$ est une matrice carrée diagonale de format(k, k) dont les éléments diagonaux sont les *valeurs singulières* de la matrice $\boldsymbol{A}$. Si $\boldsymbol{A}$ est de rang r, il y a r valeurs singulières non nulles. La matrice $\boldsymbol{U}$ est de format (n, k) et ses colonnes de même indice que les valeurs singulières non nulles sont les vecteurs propres (orthonormés) de la matrice $\boldsymbol{AA}^t$. Alors que la matrice $\boldsymbol{V}$ est une matrice carré de format (k, k) dont les colonnes de même indice que les valeurs singulières non nulles sont les vecteurs propres orthonormés de $\boldsymbol{A}^t\boldsymbol{A}$. On a $\boldsymbol{U}^t\boldsymbol{U} = \boldsymbol{I}_k$ et $\boldsymbol{V}^t\boldsymbol{V} = \boldsymbol{I}_k$. Cette décomposition s'appelle *décomposition en valeurs singulières.*

   On montre facilement que la pseudo-inverse de $\boldsymbol{D}$ est une matrice carré diagonale $\boldsymbol{D}^{(-1)}$ dont les éléments diagonaux sont les inverses des valeurs singulières non nulles alors que les valeurs singulières nulles sont laissées nulles. On obtient alors $\boldsymbol{A}^{(-1)}$ de la façon suivante :

$$\boldsymbol{A}^{(-1)} = \boldsymbol{V}\boldsymbol{D}^{(-1)}\boldsymbol{U}^t \quad \text{pour} \quad \boldsymbol{A} = \boldsymbol{UDV}^t. \tag{14.63}$$

5. Si $\boldsymbol{A}$ est une matrice carré et si $\boldsymbol{P}$ est une matrice projective orthogonale sur le sous-espace $\ker \boldsymbol{A}$, c'est-à-dire si $\boldsymbol{P}$ possède les propriétés suivantes $\boldsymbol{P}^2 = \boldsymbol{P}$, $\boldsymbol{PA} = \boldsymbol{A}^t\boldsymbol{P} = 0$, alors on a la relation :

$$\boldsymbol{A}^{(-1)} = (\boldsymbol{A} + \boldsymbol{P})^{-1} - \boldsymbol{P}\,. \tag{14.64}$$

6. Si $\boldsymbol{A}$ est carrée et symétrique elle est alors diagonalisable. Soit $\boldsymbol{\Lambda} = \boldsymbol{U}^t\boldsymbol{AU}$ sa forme diagonale. La matrice pseudo-inverse de $\boldsymbol{A}$ est donnée par :

$$\boldsymbol{A}^{(-1)} = \boldsymbol{U}(\boldsymbol{\Lambda} + \boldsymbol{\Pi})^{-1}\boldsymbol{U}^t - \boldsymbol{U\Pi U}^t\,, \tag{14.65}$$

où $\boldsymbol{\Pi}$ est la forme diagonale de la matrice projective $\boldsymbol{P}$. C'est elle-même une matrice projective dont la diagonale est nécessairement formée de 0 et de 1. Les 0 correspondent aux valeurs propres de $\boldsymbol{A}$ qui ne sont pas nulles et les 1 aux valeurs propres de $\boldsymbol{A}$ qui sont nulles.

Application à la méthode des moindres carrés.

Les concepts introduits ci-dessus s'appliquent parfaitement à notre problème d'estimation. L'espace $\mathcal{K}$ est, d'après le corollaire 14.1, le noyau de $\boldsymbol{X}$ ou de $\boldsymbol{X}^t\boldsymbol{X}$. L'espace $\mathcal{I}$ est celui des vecteurs de la forme $\boldsymbol{X}^t\boldsymbol{y}$ où $\boldsymbol{y}$ est quelconque, en d'autres termes c'est l'image de l'espace des observations par l'application duale de matrice $\boldsymbol{X}^t$. L'espace des paramètres est décomposé par $\boldsymbol{X}$ en une somme de deux sous-spaces orthogonaux $P_k = \mathcal{I} \oplus \mathcal{K}$, en ce sens qu'un vecteur quelconque de P_k peut être décomposé de façon unique en une somme d'un vecteur de $\mathcal{I}$ et d'un vecteur de $\mathcal{K}$.

La matrice pseudo-inverse permet d'écrire les solutions $\widehat{\boldsymbol{\theta}}$ des équations normales dans le cas singulier (ce qui inclu aussi le cas régulier), sous les deux formes équivalentes suivantes :

$$\widehat{\boldsymbol{\theta}} = (\boldsymbol{X}^t\boldsymbol{X})^{(-1)}\boldsymbol{X}^t\boldsymbol{y} + \boldsymbol{P}\widetilde{\boldsymbol{\theta}}, \tag{14.66}$$

$$\widehat{\boldsymbol{\theta}} = \boldsymbol{X}^{(-1)}\boldsymbol{y} + \boldsymbol{P}\widetilde{\boldsymbol{\theta}} \tag{14.67}$$

où $\widetilde{\boldsymbol{\theta}}$ est un vecteur quelconque de l'espace des paramètres et $\boldsymbol{P}$ est un projecteur orthogonal, de format (k,k), sur le noyau de $\boldsymbol{X}^t\boldsymbol{X}$ ou de $\boldsymbol{X}$. En pratique, $(\boldsymbol{X}^t\boldsymbol{X})^{(-1)}$ est plus facile à calculer que $\boldsymbol{X}^{(-1)}$. La matrice pseudo-inverse de $\boldsymbol{X}^t\boldsymbol{X}$ s'exprime à l'aide de la matrice $\boldsymbol{P}$:

$$(\boldsymbol{X}^t\boldsymbol{X})^{(-1)} = (\boldsymbol{X}^t\boldsymbol{X} + \boldsymbol{P})^{-1} - \boldsymbol{P}, \tag{14.68}$$

et réciproquement on trouve les projecteurs grâce aux matrices pseudo-inverses :

$$\begin{aligned} \boldsymbol{P} &= \boldsymbol{I}_k - \boldsymbol{Q}, \\ \boldsymbol{Q} &= (\boldsymbol{X}^t\boldsymbol{X})^{(-1)}\boldsymbol{X}^t\boldsymbol{X} = \boldsymbol{X}^{(-1)}\boldsymbol{X}\,. \end{aligned} \tag{14.69}$$

Les solutions, telles qu'elles sont exprimées par les équations (14.66) et (14.67), nous disent que l'estimation de $\widehat{\boldsymbol{\theta}}$ au sens des moindres carrés peut être mise sous la forme d'une somme $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}_1 + \widehat{\boldsymbol{\theta}}_0$, où $\widehat{\boldsymbol{\theta}}_0$ est un vecteur quelconque de l'espace $\mathcal{K}$ et d'un vecteur particulier $\widehat{\boldsymbol{\theta}}_1$ qui, d'après le théorème 14.4, appartient à l'espace $\mathcal{I}$ semblable au second membre des équations normales. Nous avons vu que la compatibilité des équations normales voulait dire que $\mathcal{I} \perp \mathcal{K}$.

Fonctions à estimer. Si $\boldsymbol{\beta}$ sont les fonctions à estimer, elles ne seront définies de façon unique, dans le cas singulier et pour des fonctions linéaires : $\boldsymbol{\beta} = \boldsymbol{C\theta}$, que si $\boldsymbol{CP} = 0$. Dans ce cas on a :

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{C}\widehat{\boldsymbol{\theta}} = \boldsymbol{C}(\boldsymbol{X}^t\boldsymbol{X} + \boldsymbol{P})^{-1}\boldsymbol{X}^t\boldsymbol{y}\,. \tag{14.70}$$

Nous souhaitons caractériser toutes les combinaisons linéaires, $\boldsymbol{\beta} = \boldsymbol{C\theta}$ des θ_i qui sont des fonctions à estimer. Nous disposons pour cela du théorème suivant :

Théorème 14.5. *Le paramètre $\beta = \boldsymbol{c}^t\boldsymbol{\theta}$ est estimée de façon unique par la méthode des moindres carrés si, et seulement si, $\boldsymbol{c} \in \mathcal{I}$.*

Démonstration. Si $\boldsymbol{c} \in \mathcal{I}$ on a par définition $\boldsymbol{Qc} = \boldsymbol{c}$ et $\boldsymbol{c}^t = \boldsymbol{c}^t\boldsymbol{Q}$ car $\boldsymbol{Q}$ est symétrique. Posons $\boldsymbol{C} = \boldsymbol{c}^t$, d'où $\boldsymbol{CP} = \boldsymbol{CQP} = \boldsymbol{0}$ car $\boldsymbol{QP} = \boldsymbol{0}$. La matrice $\boldsymbol{C}$ (formée d'une seule ligne) répond bien à la condition $\boldsymbol{CP} = \boldsymbol{0}$ et $\boldsymbol{c}^t\boldsymbol{\theta}$ est une fonction à estimer.

Réciproquement si $\boldsymbol{CP} = \boldsymbol{c}^t\boldsymbol{P} = \boldsymbol{0}$ alors, puisque $\boldsymbol{P}$ est symétrique $\boldsymbol{Pc}^t = \boldsymbol{0}$. Mais par définition $\boldsymbol{I}_k = \boldsymbol{P} + \boldsymbol{Q}$ et $(\boldsymbol{P} + \boldsymbol{Q})\boldsymbol{c} = \boldsymbol{c} = \boldsymbol{Qc}$ et $\boldsymbol{c} \in \mathcal{I}$. □

Nous savons maintenant à quoi correspond le sous-espace $\mathcal{I}$, c'est l'ensemble des combinaisons linéaires des θ_i qui peuvent être estimées sans ambiguïté par la méthode des moindres carrés.

L'espace $\mathcal{I}$ est de dimension $r = \operatorname{rg}\boldsymbol{X}$, il s'ensuit qu'au plus r combinaisons $\beta_j = \boldsymbol{c}_j^t\boldsymbol{\theta}$ linéairement indépendantes peuvent être estimées de façon unique. Par conséquent, la matrice $\boldsymbol{C}$ comporte au plus r lignes. De plus les vecteurs $\boldsymbol{c} \in \mathcal{I}$ sont, par définition, de la forme $\boldsymbol{c} = \boldsymbol{X}^t\widetilde{\boldsymbol{y}}$, où $\widetilde{\boldsymbol{y}}$ est un vecteur quelconque de O_n. Si on choisit un vecteur de la base canonique, $\boldsymbol{c}^t$ est alors égal à une ligne de $\boldsymbol{X}$ et une condition suffisante pour que r combinaisons β_j soient r fonctions à estimer linéairement indépendantes est que :

$$\beta_j = \sum_{i=1}^{k} x_{ji}\theta_i, \quad j = j_1, \ldots, j_r\,, \tag{14.71}$$

où les j correspondent aux indices de r lignes linéairement indépendantes de $\boldsymbol{X}$. Les combinaisons linéaires de ces r lignes engendrent toutes les fonctions à estimer de type $\boldsymbol{\beta} = \boldsymbol{C\theta}$.

▶**Exemple 14.5.** Nous traitons ici, à l'aide de la matrice pseudo-inverse, d'un exemple proposé par Kendall & Stuart [41] ch.19.4. Supposons que nous cherchions à estimer $k = 3$ paramètres $\boldsymbol{\theta}$ à partir de $n = 4$ observations suivant un modèle linéaire où :

$$\boldsymbol{X} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}, \quad \boldsymbol{\theta} = \begin{pmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \end{pmatrix}.$$

Le modèle est visiblement singulier de rang $r = 2$, par conséquent la matrice $\boldsymbol{X}^t\boldsymbol{X}$ est aussi de rang 2 et son noyau est de dimension 1. Il se trouve que :

$$\boldsymbol{X}^t\boldsymbol{X} = \begin{pmatrix} 4 & 2 & 2 \\ 2 & 2 & 0 \\ 2 & 0 & 2 \end{pmatrix} = \begin{pmatrix} 2 & 0 \\ 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} 2 & 1 & 1 \\ 0 & 1 & -1 \end{pmatrix}.$$

A partir de cette décomposition squelettique de $\boldsymbol{X}^t\boldsymbol{X}$ on obtient :

$$(\boldsymbol{X}^t\boldsymbol{X})^{(-1)} = \begin{pmatrix} 2 & 0 \\ 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} \frac{1}{6} & 0 \\ 0 & \frac{1}{2} \end{pmatrix} \begin{pmatrix} \frac{1}{6} & 0 \\ 0 & \frac{1}{2} \end{pmatrix} \begin{pmatrix} 2 & 1 & 1 \\ 0 & 1 & -1 \end{pmatrix} = \frac{1}{18} \begin{pmatrix} 2 & 1 & 1 \\ 1 & 5 & -4 \\ 1 & -4 & 5 \end{pmatrix},$$

et d'après (14.69) on trouve le projecteur $\boldsymbol{P}$ qui exprime la dépendance linéaire des colonnes de $\boldsymbol{X}$:

$$\boldsymbol{P} = \frac{1}{3} \begin{pmatrix} 1 & -1 & -1 \\ -1 & 1 & 1 \\ -1 & 1 & 1 \end{pmatrix} = \frac{1}{3} \begin{pmatrix} 1 \\ -1 \\ -1 \end{pmatrix} \begin{pmatrix} 1 & -1 & -1 \end{pmatrix} = \boldsymbol{z}\boldsymbol{z}^t, \quad \boldsymbol{z} \in \ker \boldsymbol{X},$$

d'où la solution générale $\widehat{\boldsymbol{\theta}} = (\boldsymbol{X}^t\boldsymbol{X})^{(-1)}\boldsymbol{X}^t\boldsymbol{y} + \boldsymbol{P}\boldsymbol{h}$:

$$\begin{pmatrix} \widehat{\theta}_1 \\ \widehat{\theta}_2 \\ \widehat{\theta}_3 \end{pmatrix} = \frac{1}{6} \begin{pmatrix} 1 & 1 & 1 & 1 \\ 2 & -1 & 2 & -1 \\ -1 & 2 & -1 & 2 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} + \frac{1}{3} \begin{pmatrix} 1 & -1 & -1 \\ -1 & 1 & 1 \\ -1 & 1 & 1 \end{pmatrix} \begin{pmatrix} h_1 \\ h_2 \\ h_3 \end{pmatrix},$$

où h_1, h_2 et h_3 sont des nombres quelconques. Pour les fonctions à estimer (ou plutôt : qu'il est possible d'estimer), les deux premières lignes de $\boldsymbol{X}$ sont linéairement indépendantes, ce qui donne une première estimation possible :

$$\begin{pmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} \widehat{\theta}_1 \\ \widehat{\theta}_2 \\ \widehat{\theta}_3 \end{pmatrix} = \begin{pmatrix} \widehat{\theta}_1 + \widehat{\theta}_2 \\ \widehat{\theta}_1 + \widehat{\theta}_3 \end{pmatrix} = \frac{1}{2} \begin{pmatrix} y_1 + y_3 \\ y_2 + y_4 \end{pmatrix}.$$

Mais toute combinaison linéaire des lignes de $\boldsymbol{C}$ est aussi une fonction à estimer, par exemple :

$$\begin{pmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{pmatrix} = \begin{pmatrix} 2 & 1 & 1 \\ 0 & 1 & -1 \end{pmatrix} \begin{pmatrix} \widehat{\theta}_1 \\ \widehat{\theta}_2 \\ \widehat{\theta}_3 \end{pmatrix} = \begin{pmatrix} 2\widehat{\theta}_1 + \widehat{\theta}_2 + \widehat{\theta}_3 \\ \widehat{\theta}_2 - \widehat{\theta}_3 \end{pmatrix} = \frac{1}{2} \begin{pmatrix} y_1 + y_2 + y_3 + y_4 \\ y_1 - y_2 + y_3 - y_4 \end{pmatrix}.$$

En revanche, il n'est pas possible d'estimer (de façon linéaire et non ambiguë) un paramètre unique quelconque θ_i, car $\boldsymbol{I}_3\boldsymbol{P} = \boldsymbol{P}$ et aucune ligne de ce projecteur n'est nulle. Donnons pour terminer la matrice des variances-covariances de $\widehat{\boldsymbol{\beta}}$ qui, nous le montrerons, s'obtient par (14.93) c'est-à-dire par $\boldsymbol{V}_{\widehat{\beta}} = \sigma^2\boldsymbol{C}(\boldsymbol{X}^t\boldsymbol{X})^{(-1)}\boldsymbol{C}^t$:

$$\boldsymbol{V}_{\widehat{\beta}} = \sigma^2 \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{pmatrix} \quad \text{pour les premières,}$$

$$\boldsymbol{V}_{\widehat{\beta}} = \sigma^2 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \text{pour les secondes.}$$

14.3.6 Reparamétrisation du modèle.

Nous avons déjà évoqué la reparamétrisation du modèle dans le cas régulier lorsque nous avons cherché des bases orthogonales. Nous n'envisagerons donc ici que le cas singulier.

Nous savons qu'une condition nécessaire et suffisante pour que le modèle soit singulier est que les colonnes de la matrice $\boldsymbol{X}$ soient linéairement dépendantes. Cette circonstance nous indique que le modèle choisi, et par conséquent sa paramétrisation en $\boldsymbol{\theta}$, n'est pas adaptée. Nous voulons maintenant montrer qu'il est toujours possible de choisir une autre paramétrisation $\boldsymbol{\beta} = \boldsymbol{C}\boldsymbol{\theta}$ où la matrice modèle $\boldsymbol{X}_\beta$ qui en découle n'est pas singulière.

Remarquons qu'une décomposition squelettique quelconque de $\boldsymbol{X}$ réalise une nouvelle paramétrisation régulière en $\boldsymbol{\beta}$. En effet on a $\boldsymbol{X} = \boldsymbol{X}_\beta \boldsymbol{C}$, où $\boldsymbol{X}_\beta$ et $\boldsymbol{C}$ sont de rang r, et $\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\theta} = \boldsymbol{X}_\beta \boldsymbol{C}\boldsymbol{\theta} = \boldsymbol{X}_\beta \boldsymbol{\beta}$. Les matrices $\boldsymbol{C}\boldsymbol{C}^t$ et $\boldsymbol{X}_\beta^t \boldsymbol{X}_\beta$ sont inversibles et il est facile de montrer à partir de la relation (14.62) que $\boldsymbol{Q} = \boldsymbol{C}^t (\boldsymbol{C}\boldsymbol{C}^t)^{-1} \boldsymbol{C}$, $\boldsymbol{P} = \boldsymbol{I}_k - \boldsymbol{Q}$ et de là que $\boldsymbol{C}\boldsymbol{P} = \boldsymbol{0}$.

Il reste à trouver, $\boldsymbol{C}$ étant donnée, comment calculer le modéle $\boldsymbol{X}_\beta$ correspondant. Il faut résoudre $\boldsymbol{X} = \boldsymbol{X}_\beta \boldsymbol{C}$ qui représente n systèmes linéaires de k équations à $r \leq k$ inconnues (les r éléments des n lignes de $\boldsymbol{X}_\beta$.) Ces équations sont compatibles, en effet la condition de Fredholm est satisfaite si $\boldsymbol{C}\boldsymbol{z} = \boldsymbol{0} \Rightarrow \boldsymbol{X}\boldsymbol{z} = \boldsymbol{0}$. Les lignes de $\boldsymbol{C}$ forment une base de $\mathcal{I}$ $\boldsymbol{z}$ appartient donc au sous-espace orthogonal à $\mathcal{I}$ c'est-à-dire à $\mathcal{K}$ qui est $\ker \boldsymbol{X}$ par définition, on a donc $\boldsymbol{X}\boldsymbol{z} = \boldsymbol{0}$.

On peut alors résoudre le système en supprimant $k - r$ équations, mais on peut aussi donner la solution en termes de pseudo-inverse, il vient

$$\boldsymbol{X}_\beta = \boldsymbol{X}\boldsymbol{C}^{(-1)} = \boldsymbol{X}\boldsymbol{C}^t (\boldsymbol{C}\boldsymbol{C}^t)^{-1}. \tag{14.72}$$

De $\boldsymbol{X} = \boldsymbol{X}_\beta \boldsymbol{C}$ on tire $\operatorname{rg} \boldsymbol{X} \leq \operatorname{rg} \boldsymbol{X}_\beta$ et de (14.72) on tire $\operatorname{rg} \boldsymbol{X}_\beta \leq \operatorname{rg} \boldsymbol{X}$, d'où $\operatorname{rg} \boldsymbol{X}_\beta = \operatorname{rg} \boldsymbol{X} = r$. La matrice $\boldsymbol{X}_\beta^t \boldsymbol{X}_\beta$ est alors régulière et les r paramètres $\boldsymbol{\beta}$ sont estimés de façon unique par :

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}_\beta^t \boldsymbol{X}_\beta)^{-1} \boldsymbol{X}_\beta^t \boldsymbol{y}. \tag{14.73}$$

Cette reparamétrisation peut toujours être faite. On postule d'habitude qu'elle a bien été faite et on écrit alors le modèle linéaire sous la forme $\boldsymbol{y} = \boldsymbol{X}_\beta \boldsymbol{\beta} + \boldsymbol{\epsilon}$. Cette procédure est justifiée tant que c'est le signal $\boldsymbol{\mu}$ qui est la quantité dont l'estimation importe et que les paramètres ($\boldsymbol{\theta}$ ou $\boldsymbol{\beta}$) ne servent que d'intermédiaires pour atteindre ce but. Cependant il peut exister des cas où l'estimation de $\boldsymbol{\theta}$ est aussi requise, la reparamétrisation en $\boldsymbol{\beta}$ nous donne alors une base de l'espace des seules combinaisons linéaires de $\boldsymbol{\theta}$ qui peuvent être estimées de façon unique par la méthode des moindres carrés.

▶**Exemple 14.6.** En reprenant l'exemple précédent on trouve pour la première série de paramètres, soit par calcul direct soit en utilisant (14.72) :

$$\boldsymbol{X} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}, \quad \boldsymbol{C} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}, \quad \boldsymbol{X}_\beta = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix},$$

alors que l'on trouve pour la seconde :

$$\boldsymbol{C} = \begin{pmatrix} 2 & 1 & 1 \\ 0 & 1 & -1 \end{pmatrix}, \quad \boldsymbol{X}_\beta = \frac{1}{2} \begin{pmatrix} 1 & 1 \\ 1 & -1 \\ 1 & 1 \\ 1 & -1 \end{pmatrix}.$$

14.3.7 Interprétation géométrique de la méthode des moindres carrés, dans l'espace des observations.

Rappelons tout d'abord les données du problème. Le vecteur $\boldsymbol{y}$ des observations est formé de n valeurs connues y_i observées aux points x_i. Ces valeurs sont entachées d'une erreur ϵ_i de valeur inconnue telle que $y_i = \mu_i + \epsilon_i$, les ϵ_i sont

les composantes d'un vecteur de bruit $\boldsymbol{\epsilon}$. L'observation $\boldsymbol{y}$ et le bruit $\boldsymbol{\epsilon}$ sont des vecteurs colonnes susceptibles d'appartenir à tout l'espace arithmétique : O_n, dit « espace des observations ». Sous hypothèse qu'il est possible de réduire les variables aléatoires ϵ_i on a $\mathrm{E}\{\epsilon_i\} = 0$ et $\mathrm{E}\{\epsilon_i \epsilon_j\} = \sigma^2 \delta_{ij}$. Les valeurs moyennes des observations forment ce qu'on nomme un signal $\boldsymbol{\mu}$ qu'il est, par hypothèse, possible de représenter linéairement à l'aide des k paramètres θ_j, suivant le modèle $\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\theta}$. Le vecteur des paramètres $\boldsymbol{\theta}$ est susceptible d'appartenir à tout l'espace arithmétique P_k. La matrice du modèle : $\boldsymbol{X}$ est de rang $\mathrm{rg}\,\boldsymbol{X} = r \leq k$ et on suppose qu'une reparamétrisation quelconque $\boldsymbol{\mu} = \boldsymbol{X}_\beta \boldsymbol{\beta}$ a été faite de façon à ce que la matrice $\boldsymbol{X}_\beta$ soit de rang r. Cette matrice définit un sous-espace de dimension r : $M_r \subseteq O_n$ où se trouve $\boldsymbol{\mu}$.

Afin d'alléger la notation nous supprimons l'indice $_\beta$ de la matrice $\boldsymbol{X}_\beta$. Nous notons $\widetilde{\boldsymbol{\beta}}$ une estimation quelconque de $\boldsymbol{\beta}$ et les estimateurs des moindres carrés sont notés $\widehat{\boldsymbol{\beta}}$ et $\widehat{\boldsymbol{\mu}}$. Cherchons à présent à dégager les relations métriques qui existent entre toutes les quantités que nous avons introduites. On pourra s'aider de la figure 14.2 pour visualiser les résultats.

Relations entre l'observation y et l'estimation $\widehat{\mu}$ de la moyenne.

Nous avons introduit une estimation de $\boldsymbol{\mu}$ obtenue par l'équation $\widehat{\boldsymbol{\mu}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$. Remplaçons $\widehat{\boldsymbol{\beta}}$ par sa valeur, il vient $\widehat{\boldsymbol{\mu}} = \boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{y}$. L'estimation $\widehat{\boldsymbol{\mu}}$ est trouvée comme combinaison linéaire de l'observation $\boldsymbol{y}$. Soit $\boldsymbol{H}$ la matrice de cette combinaison linéaire, on a alors :

$$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t, \qquad \widehat{\boldsymbol{\mu}} = \boldsymbol{H}\boldsymbol{y}. \tag{14.74}$$

On appelle parfois $\boldsymbol{H}$ la matrice « *chapeau* », car son rôle est de construire une estimation, de « mettre un chapeau » en quelque sorte. Etablissons maintenant quelques propriétés de $\boldsymbol{H}$.

1. $\boldsymbol{H}$ est une matrice (n, n) symétrique : $\boldsymbol{H} = \boldsymbol{H}^t$, ce fait est établi facilement. En tant que matrice symétrique, $\boldsymbol{H}$ est diagonalisable.

2. $\boldsymbol{H}$ est un projecteur, c'est-à-dire que $\boldsymbol{H}^2 = \boldsymbol{H}$. En effet, $\boldsymbol{H}^2 = \boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t = \boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t = \boldsymbol{H}$. En tant que projecteur $\boldsymbol{H}$ a pour valeurs propres 0 ou 1, et son rang est égal à sa trace. Son rang est égal à la dimension du sous-espace image sur lequel il projette ; c'est-à-dire qu'il est égal à la multiplicité de la valeur propre 1. Il suffit de considérer la forme diagonale de $\boldsymbol{H}$ pour s'en convaincre.

3. La trace de $\boldsymbol{H}$ est égale à $r = \mathrm{rg}\,\boldsymbol{X}$. Ce fait est trivial puisque d'après (14.74) $\boldsymbol{H}$ projette $\boldsymbol{y}$ sur un sous-espace de dimension k. Donnons-en, cependant, une démonstration directe. On a, $\mathrm{trace}(\boldsymbol{H}) = \mathrm{trace}(\boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t) = \mathrm{trace}(\boldsymbol{X}^t\boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X})^{-1}) = \mathrm{trace}(\boldsymbol{I}_r) = r$, où la matrice $\boldsymbol{I}_r$ désigne la matrice identité de format (r, r). Dans les manipulations sous l'opérateur trace, on peut changer l'ordre des multiplications matricielles tout en prenant garde à ce qu'elles restent cohérentes car $\boldsymbol{X}$ est une matrice rectangulaire.

4. On a les relations $\boldsymbol{H}\widehat{\boldsymbol{\mu}} = \widehat{\boldsymbol{\mu}}$ et $\boldsymbol{H}\boldsymbol{\mu} = \boldsymbol{\mu}$. Ces relations sont également évidentes car, par définition, $\widehat{\boldsymbol{\mu}}$ et $\boldsymbol{\mu}$ appartient à l'espace image de $\boldsymbol{H}$.

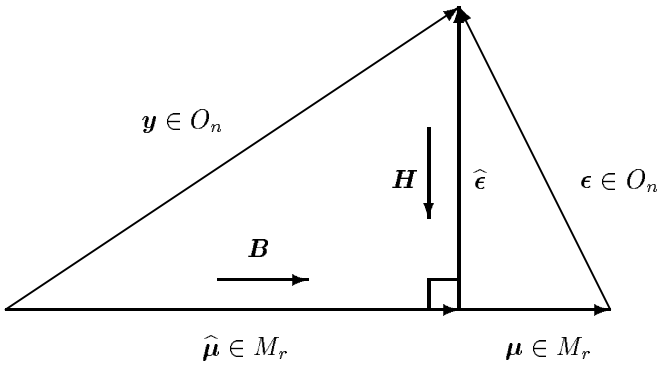


FIG. 14.2: *Interprétation géométrique de la méthode des moindres carrés, et mise en évidence de la relation d'orthogonalité. Les opérateurs* $\boldsymbol{H}$ *et* $\boldsymbol{B}$ *sont des projecteurs. Les vecteurs* $\boldsymbol{\mu}$ *et* $\widehat{\boldsymbol{\mu}}$*, ici colinéaires, appartiennent au même sous-espace* M_r*, de dimension égale au rang de la matrice du modèle* $r = \mathrm{rg}\,\boldsymbol{X}$*. Naturellement si* $r > 1$*, ces vecteurs ne sont pas nécessairement colinéaires.*

Relations entre l'observation y et le résidu $\widehat{\epsilon}$.

On a par définition, $\widehat{\boldsymbol{\epsilon}} = \boldsymbol{y} - \widehat{\boldsymbol{\mu}}$, soit $\widehat{\boldsymbol{\epsilon}} = \boldsymbol{y} - \boldsymbol{H}\boldsymbol{y} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}$. Le résidu $\widehat{\boldsymbol{\epsilon}}$ est également trouvé comme combinaison linéaire de l'observation $\boldsymbol{y}$. Soit $\boldsymbol{B}$ la matrice de la combinaison linéaire correspondante. On a $\boldsymbol{B} = \boldsymbol{I} - \boldsymbol{H}$, et donc :

$$\widehat{\boldsymbol{\epsilon}} = \boldsymbol{B}\boldsymbol{y}\,. \tag{14.75}$$

Tournons-nous à présent vers les propriétés de $\boldsymbol{B}$ et donnons, en complément, certaines relations concernant $\boldsymbol{H}$ et $\boldsymbol{B}$.

1. $\boldsymbol{B}$ est symétrique : ce fait découle de la propriété correspondante de $\boldsymbol{H}$. Comme $\boldsymbol{H}$, $\boldsymbol{B}$ est alors diagonalisable.

2. $\boldsymbol{B}$ est un projecteur : $\boldsymbol{B}^2 = \boldsymbol{B}$. En effet, $\boldsymbol{B}^2 = (\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{I} - \boldsymbol{H}) = \boldsymbol{I} - 2\boldsymbol{H} + \boldsymbol{H}^2 = \boldsymbol{I} - \boldsymbol{H} = \boldsymbol{B}$.

3. La trace de $\boldsymbol{B}$ est égale à $n - r$. En effet, $\mathrm{trace}(\boldsymbol{B}) = \mathrm{trace}(\boldsymbol{I} - \boldsymbol{H}) = \mathrm{trace}(\boldsymbol{I}) - \mathrm{trace}(\boldsymbol{H}) = n - r$, $r = \mathrm{rg}\,\boldsymbol{X}$.

4. Les projecteurs $\boldsymbol{H}$ et $\boldsymbol{B}$, sont orthogonaux. Pour cela il faut vérifier que $\boldsymbol{H}\boldsymbol{B} = \boldsymbol{B}\boldsymbol{H} = 0$, ce que l'on démontre facilement : $\boldsymbol{H}\boldsymbol{B} = \boldsymbol{H}(\boldsymbol{I} - \boldsymbol{H}) = \boldsymbol{H} - \boldsymbol{H} = 0$, et de même pour $\boldsymbol{B}\boldsymbol{H}$.

5. On a les relations $\boldsymbol{B}\widehat{\boldsymbol{\epsilon}} = \widehat{\boldsymbol{\epsilon}}$ et $\boldsymbol{B}\boldsymbol{\epsilon} = \widehat{\boldsymbol{\epsilon}}$. La première relation est triviale car, par définition, $\boldsymbol{\epsilon}$ appartient à l'espace image de $\boldsymbol{B}$. Pour la deuxième, on a $\boldsymbol{B}\boldsymbol{\epsilon} = \boldsymbol{B}(\boldsymbol{y} - \boldsymbol{\mu}) = \widehat{\boldsymbol{\epsilon}} - \boldsymbol{B}\boldsymbol{\mu} = \widehat{\boldsymbol{\epsilon}} - \boldsymbol{B}\boldsymbol{H}\boldsymbol{\mu} = \widehat{\boldsymbol{\epsilon}}$.

6. Finalement on a, $\boldsymbol{B}\boldsymbol{\mu} = \boldsymbol{B}\widehat{\boldsymbol{\mu}} = \boldsymbol{0}$ et $\boldsymbol{H}\widehat{\boldsymbol{\epsilon}} = \boldsymbol{0}$, ce que l'on démontre facilement à l'aide de la relation $\boldsymbol{H}\boldsymbol{B} = \boldsymbol{B}\boldsymbol{H} = \boldsymbol{0}$.

Relation d'orthogonalité.

Etablissons maintenant que le vecteur $\widehat{\boldsymbol{\epsilon}}$ est orthogonal à $\widehat{\boldsymbol{\mu}}$. Il vient :

$$(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\epsilon}}) = \widehat{\boldsymbol{\mu}}^t \widehat{\boldsymbol{\epsilon}} = \widehat{\boldsymbol{\mu}}^t \boldsymbol{B} \widehat{\boldsymbol{\epsilon}} = (\boldsymbol{B}\widehat{\boldsymbol{\mu}})^t \widehat{\boldsymbol{\epsilon}} = \boldsymbol{0}, \tag{14.76}$$

et ces deux vecteurs sont bien orthogonaux. Cette propriété correspond au principe d'orthogonalité de la théorie de l'estimation linéaire que nous avons trouvée ici comme conséquence de la méthode des moindres carrés. Si $\widetilde{\boldsymbol{\mu}}$ est un vecteur quelconque du sous-espace image de $\boldsymbol{H}$, $\widetilde{\boldsymbol{y}} \in \mathrm{ima}(\boldsymbol{H}) = M_r$, on a de la même façon $(\widetilde{\boldsymbol{\mu}}, \widehat{\boldsymbol{\epsilon}}) = \widetilde{\boldsymbol{\mu}}^t \widehat{\boldsymbol{\epsilon}} = \widetilde{\boldsymbol{\mu}}^t \boldsymbol{H}^t \widehat{\boldsymbol{\epsilon}} = \widetilde{\boldsymbol{\mu}}^t \boldsymbol{H} \widehat{\boldsymbol{\epsilon}} = 0$, soit donc :

$$(\widetilde{\boldsymbol{\mu}}, \widehat{\boldsymbol{\epsilon}}) = 0, \quad \widetilde{\boldsymbol{\mu}} \in M_r \,. \tag{14.77}$$

Le résidu $\widehat{\boldsymbol{\epsilon}}$ est donc orthogonal au sous-espace M_r auquel appartiennent, en particulier, $\boldsymbol{\mu}, \widehat{\boldsymbol{\mu}}, \boldsymbol{\theta}$ et $\widehat{\boldsymbol{\theta}}$. On peut dire alors que l'estimation $\widehat{\boldsymbol{\mu}}$ est la projection orthogonale de l'observation $\boldsymbol{y}$ sur le sous-espace de tous les modèles possibles $\widetilde{\boldsymbol{\mu}}$. Ces propriétés sont également illustrées par la figure 14.2.

Moindres carrés pondérés.

Dans le cas où l'on introduit la matrice de pondération relative $\boldsymbol{V}^{-1}$, le changement de base de matrice $(\boldsymbol{N}^t)^{-1}$ nous a montré que l'on pouvait se ramener au cas précédent, à la condition naturellement que $\boldsymbol{N}$ soit non-singulière. Les projecteurs prennent alors la forme suivante :

$$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^t \boldsymbol{V}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^t \boldsymbol{V}^{-1} \,, \qquad \boldsymbol{B} = \boldsymbol{I} - \boldsymbol{H}, \tag{14.78}$$

et le produit scalaire devient :

$$(\boldsymbol{y}_1, \boldsymbol{y}_2) = \boldsymbol{y}_1^t \boldsymbol{V}^{-1} \boldsymbol{y}_2 \,. \tag{14.79}$$

14.3.8 Le théorème de Gauss-Markov dans le cas linéaire de la méthode des moindres carrés.

Le théorème de Gauss-Markov établit les propriétés optimales de l'estimation d'une combinaison linéaire $\boldsymbol{C\theta}$ des paramètres, calculée à partir de l'estimateur $\widehat{\boldsymbol{\theta}}$. Faisons trois remarques préliminaires à propos de la matrice $\boldsymbol{C}$.

- Cette matrice est supposée connue, mais elle n'est pas nécessairement de rang k. Elle peut être de rang $r \leq k$ si nous ne nous intéressons qu'à un nombre restreint r de combinaisons linéaires des θ_i.

- Dans le cas régulier, $\boldsymbol{C}$ peut être quelconque, mais dans le cas singulier il faut que $\boldsymbol{CP} = \boldsymbol{0}$, où $\boldsymbol{P}$ est un projecteur orthogonal sur le noyau de $\boldsymbol{X}^t\boldsymbol{X}$ donné par l'équation (14.69).

- Les résultats du théorème de Gauss-Markov s'appliquent naturellement dans le cas régulier à l'estimateur des moindres carrés $\widehat{\boldsymbol{\theta}}$ pour la valeur particulière de $\boldsymbol{C}$ égale à l'identité.

Ces remarques étant faites, passons maintenant à l'énoncé du théorème.

Théorème 14.6. *(Gauss-Markov.) Dans le modèle linéaire $\boldsymbol{y} = \boldsymbol{X\theta} + \boldsymbol{\epsilon}$, où le bruit $\boldsymbol{\epsilon}$ suit une loi quelconque de moyenne nulle et de matrice des variances--covariances finie, l'estimateur $\boldsymbol{C\widehat{\theta}}$, où $\widehat{\boldsymbol{\theta}}$ est l'estimateur des moindres carrés de $\boldsymbol{\theta}$, possède les propriétés suivantes :*

1. *L'estimateur $\widehat{\boldsymbol{\beta}} = \boldsymbol{C\widehat{\theta}}$ est non-biaisé pour l'estimation de $\boldsymbol{\beta} = \boldsymbol{C\theta}$.*
2. *Il a la plus petite variance dans la classe des estimateurs de $\boldsymbol{C\theta}$ non-biaisés et obtenus comme combinaison linéaire de l'observation $\boldsymbol{y}$.*
3. *Les vecteurs aléatoires, $\boldsymbol{C\widehat{\theta}}$ et $\widehat{\boldsymbol{\epsilon}} = \boldsymbol{y} - \boldsymbol{X\widehat{\theta}}$ sont non-corrélés.*

Démonstration. 1) Soit $\boldsymbol{t}$ un estimateur linéaire quelconque de $\boldsymbol{C\theta}$. On définit $\boldsymbol{T}$ comme étant la matrice permettant de calculer l'estimateur à partir des observations. On a $\boldsymbol{t} = \boldsymbol{Ty}$, soit $\boldsymbol{t} = \boldsymbol{TX\theta} + \boldsymbol{T\epsilon}$. La condition pour que $\boldsymbol{t}$ soit non-biaisé est $\mathrm{E}\{\boldsymbol{t}\} = \boldsymbol{C\theta}$, c'est-à-dire : $\mathrm{E}\{\boldsymbol{t}\} = \mathrm{E}\{\boldsymbol{Ty}\} = \mathrm{E}\{\boldsymbol{T}(\boldsymbol{X\theta} + \boldsymbol{\epsilon})\} = \boldsymbol{TX\theta}$. Pour que $\boldsymbol{t}$ soit non-biaisé, il faut que $\mathrm{E}\{t\} = \boldsymbol{C\theta}$ et donc que la matrice $\boldsymbol{T}$ satisfasse l'équation :

$$\boldsymbol{TX} = \boldsymbol{C}\,. \tag{14.80}$$

Dorénavant, nous ne nous intéresserons qu'à des transformations $\boldsymbol{T}$ possédant cette propriété. Considérons l'estimateur $\widehat{\boldsymbol{t}} = \boldsymbol{C\widehat{\theta}}$, où $\widehat{\boldsymbol{\theta}}$ est l'estimateur des moindres carrés donné par l'équation (14.66). Soit $\boldsymbol{T}_0$ la matrice de la combinaison linéaire de $\boldsymbol{y}$ correspondant à cet estimateur. On a, à la condition que $\boldsymbol{CP} = 0$:

$$\boldsymbol{T}_0 = \boldsymbol{C}(\boldsymbol{X}^t\boldsymbol{X})^{(-1)}\boldsymbol{X}^t\,. \tag{14.81}$$

Calculons $\boldsymbol{T}_0\boldsymbol{X}$. Il vient en utilisant (14.54) :

$$\boldsymbol{T}_0\boldsymbol{X} = \boldsymbol{C}(\boldsymbol{X}^t\boldsymbol{X})^{(-1)}\boldsymbol{X}^t\boldsymbol{X} = \boldsymbol{C}(\boldsymbol{I}_k - \boldsymbol{P}) = \boldsymbol{C}\,. \tag{14.82}$$

L'estimateur $\widehat{\boldsymbol{t}}$ répond à la condition (14.80) et il est donc non-biaisé.

2) Vérifions maintenant que $\widehat{\boldsymbol{t}}$ est MV, c'est-à-dire que les éléments diagonaux de sa matrice des variances-covariances sont plus petits que les éléments correspondants de la matrice des variances-covariances de tout autre estimateur $\boldsymbol{t}$. Pour mener à bien ce calcul, évaluons au préalable la covariance de deux estimateurs linéaires non-biaisés quelconques $\boldsymbol{t}_1$ et $\boldsymbol{t}_2$ de matrice $\boldsymbol{T}_1$ et $\boldsymbol{T}_2$. On a, $\mathrm{Cov}(\boldsymbol{t}_1, \boldsymbol{t}_2) = \mathrm{E}\{(\boldsymbol{t}_1 - \boldsymbol{C\theta})(\boldsymbol{t}_2 - \boldsymbol{C\theta})^t\}$, mais $\boldsymbol{t} = \boldsymbol{Ty} = \boldsymbol{TX\theta} + \boldsymbol{T\epsilon}$ et d'après la condition (14.80), il vient $\boldsymbol{t} = \boldsymbol{C\theta} + \boldsymbol{T\epsilon}$, d'où $\boldsymbol{t} - \boldsymbol{C\theta} = \boldsymbol{T\epsilon}$, soit :

$$\mathrm{E}\{(\boldsymbol{t}_1 - \boldsymbol{C\theta})(\boldsymbol{t}_2 - \boldsymbol{C\theta})^t\} = \mathrm{E}\{\boldsymbol{T}_1\boldsymbol{\epsilon}\boldsymbol{\epsilon}^t\boldsymbol{T}_2^t\} = \sigma^2\boldsymbol{T}_1\boldsymbol{T}_2^t\,. \tag{14.83}$$

On a donc finalement :

$$\mathrm{Cov}(\boldsymbol{t}_1, \boldsymbol{t}_2) = \sigma^2\boldsymbol{T}_1\boldsymbol{T}_2^t\,. \tag{14.84}$$

Calculons à présent la variance de $\boldsymbol{t}$ en fonction de la variance de $\widehat{\boldsymbol{t}}$. En posant $\boldsymbol{t} = (\boldsymbol{t} - \widehat{\boldsymbol{t}}) + \widehat{\boldsymbol{t}}$, il vient :

$$\boldsymbol{V}(\boldsymbol{t}) = \boldsymbol{V}(\boldsymbol{t} - \widehat{\boldsymbol{t}}) + \boldsymbol{V}(\widehat{\boldsymbol{t}}) + \mathrm{Cov}(\boldsymbol{t} - \widehat{\boldsymbol{t}}, \widehat{\boldsymbol{t}}) + \mathrm{Cov}(\widehat{\boldsymbol{t}}, \boldsymbol{t} - \widehat{\boldsymbol{t}})\,. \tag{14.85}$$

Mais $\mathrm{Cov}(\boldsymbol{t} - \widehat{\boldsymbol{t}}, \widehat{\boldsymbol{t}}) = \mathrm{Cov}(\boldsymbol{t}, \widehat{\boldsymbol{t}}) - \boldsymbol{V}(\widehat{\boldsymbol{t}})$, évaluons ces termes grâce à l'équation (14.84). On trouve alors :

$$\begin{aligned}
\mathrm{Cov}(\boldsymbol{t}, \widehat{\boldsymbol{t}}) - \boldsymbol{V}(\widehat{\boldsymbol{t}}) &= \sigma^2\boldsymbol{T}\boldsymbol{T}_0^t - \sigma^2\boldsymbol{T}_0\boldsymbol{T}_0^t = \sigma^2(\boldsymbol{T} - \boldsymbol{T}_0)\boldsymbol{T}_0^t \\
&= \sigma^2(\boldsymbol{T} - \boldsymbol{T}_0)\boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X})^{(-1)}\boldsymbol{C}^t \\
&= \sigma^2(\boldsymbol{TX} - \boldsymbol{T}_0\boldsymbol{X})(\boldsymbol{X}^t\boldsymbol{X})^{(-1)}\boldsymbol{C}^t = 0\,.
\end{aligned}$$

Cette dernière équation est nulle car les estimateurs $\boldsymbol{t}$ et $\widehat{\boldsymbol{t}}$ étant non-biaisés, leur matrice $\boldsymbol{T}$ obéit alors à la relation (14.80), et on a donc $\boldsymbol{TX} = \boldsymbol{T}_0\boldsymbol{X} = \boldsymbol{C}$. Ainsi $\mathrm{Cov}(\boldsymbol{t} - \widehat{\boldsymbol{t}}, \widehat{\boldsymbol{t}}) = 0$ et de même pour $\mathrm{Cov}(\widehat{\boldsymbol{t}}, \boldsymbol{t} - \widehat{\boldsymbol{t}}) = 0$, et il reste dans (14.85) :

$$\boldsymbol{V}(\boldsymbol{t}) = \boldsymbol{V}(\boldsymbol{t} - \widehat{\boldsymbol{t}}) + \boldsymbol{V}(\widehat{\boldsymbol{t}})\,. \tag{14.86}$$

Les éléments diagonaux d'une matrice des variances-covariances ne sont pas négatifs car ce sont des variances. On a donc l'inégalité suivante sur les variances $[\boldsymbol{V}(\boldsymbol{t})]_{ii}$ des estimateurs t_i, éléments du vecteur colonne $\boldsymbol{t}$:

$$[\boldsymbol{V}(\boldsymbol{t})]_{ii} \geq [\boldsymbol{V}(\widehat{\boldsymbol{t}})]_{ii} \tag{14.87}$$

Donc les éléments de $\widehat{\boldsymbol{t}} = \boldsymbol{C}\widehat{\boldsymbol{\theta}}$ sont MV dans la classe des estimateurs linéaires non-biaisés de $\boldsymbol{C\theta}$. On dit que $\widehat{\boldsymbol{t}}$ est lui-même MV si tous ses éléments sont MV. Nous prouvons ainsi par la même occasion que, dans le cas régulier où $\boldsymbol{C} = \boldsymbol{I}$, l'estimateur $\widehat{\boldsymbol{\theta}}$ est également MV parmi tous les estimateurs linéaires de $\boldsymbol{\theta}$. Au cours de cette démonstration, nous avons calculé la matrice des variances-covariances $\boldsymbol{V}(\widehat{\boldsymbol{t}})$ de l'estimateur $\widehat{\boldsymbol{t}} = \boldsymbol{C}\widehat{\boldsymbol{\theta}}$, on a $\boldsymbol{V}(\widehat{\boldsymbol{t}}) = \sigma^2\boldsymbol{TT}^t$ mais $\boldsymbol{T} = \boldsymbol{C}(\boldsymbol{X}^t\boldsymbol{X})^{(-1)}\boldsymbol{X}^t$, d'où d'après (14.54) :

$$\boldsymbol{V}(\widehat{\boldsymbol{t}}) = \sigma^2\boldsymbol{C}(\boldsymbol{X}^t\boldsymbol{X})^{(-1)}\boldsymbol{C}^t\,. \tag{14.88}$$

3) Par définition $\boldsymbol{C}\widehat{\boldsymbol{\theta}}$ et $\widehat{\boldsymbol{\epsilon}}$ ne sont pas corrélés si $\mathrm{Cov}(\boldsymbol{C}\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\epsilon}}) = 0$. L'estimateur $\boldsymbol{C}\widehat{\boldsymbol{\theta}}$ est non-biaisé et $\mathrm{E}\{\boldsymbol{C}\widehat{\boldsymbol{\theta}}\} = \boldsymbol{C\theta}$. L'estimateur $\widehat{\boldsymbol{\theta}}$ est également non-biaisé et le résidu $\widehat{\boldsymbol{\epsilon}}$ est alors de moyenne nulle $\mathrm{E}\{\widehat{\boldsymbol{\epsilon}}\} = 0$. Il vient :

$$\begin{aligned}\mathrm{Cov}(\boldsymbol{C}\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\epsilon}}) &= \mathrm{E}\{(\boldsymbol{C}\widehat{\boldsymbol{\theta}} - \boldsymbol{C\theta})\widehat{\boldsymbol{\epsilon}}^t\} \\ &= \mathrm{E}\{\boldsymbol{C}\widehat{\boldsymbol{\theta}}\widehat{\boldsymbol{\epsilon}}^t\} - \boldsymbol{C\theta}\,\mathrm{E}\{\widehat{\boldsymbol{\epsilon}}^t\} \\ &= 0\,.\end{aligned}$$

Le premier terme est nul parce que $\boldsymbol{C}\widehat{\boldsymbol{\theta}} \in P_k$ et que $\widehat{\boldsymbol{\epsilon}}$ est orthogonal à P_k, et le second est également nul parce que $\widehat{\boldsymbol{\epsilon}}$ est de moyenne nulle.

Cela termine la démonstration du théorème de Gauss-Markov dans le cas des estimateurs linéaires non-biaisés d'un modèle linéaire. □

14.3.9 Moyenne et variance des estimateurs des moindres carrés.

La démonstration du théorème de Gauss-Markov nous a fourni la moyenne et la matrice des variances-covariances de la loi suivie par les estimateurs des moindres carrés. Explicitons ce résultat en distinguant une fois de plus l'estimation de $\boldsymbol{\theta}$ dans le cas régulier et l'estimation de $\boldsymbol{C\theta}$ dans le cas général.

Moyenne. Comme les estimateurs des $\widehat{\boldsymbol{\theta}}$ et $\widehat{\boldsymbol{\beta}} = \boldsymbol{C}\widehat{\boldsymbol{\theta}}$ sont non-biaisé, on a :

$$\mathrm{E}\{\widehat{\boldsymbol{\theta}}\} = \boldsymbol{\theta}, \quad \mathrm{E}\{\boldsymbol{C}\widehat{\boldsymbol{\theta}}\} = \boldsymbol{C\theta}. \tag{14.89}$$

A ces estimations non-biaisées correspond une estimation également non-biaisée de $\boldsymbol{\mu}$: $\mathrm{E}\{\widehat{\boldsymbol{\mu}}\} = \mathrm{E}\{\boldsymbol{XC}\widehat{\boldsymbol{\theta}}\} = \boldsymbol{X}\,\mathrm{E}\{\boldsymbol{C}\widehat{\boldsymbol{\theta}}\} = \boldsymbol{XC\theta} = \boldsymbol{\mu}$.

Variance des paramètres estimés dans le cas régulier. La matrice des variances-covariances $\boldsymbol{V}_{\hat{\theta}}$ des estimateurs $\widehat{\boldsymbol{\theta}}$, (à ne pas confondre avec la matrice $\boldsymbol{V}$ des variances-covariances du bruit $\boldsymbol{\epsilon}$), est donnée par l'équation (14.88). On a dans le cas réduit :

$$\boldsymbol{V}_{\hat{\theta}} = \sigma^2 (\boldsymbol{X}^t \boldsymbol{X})^{-1} , \tag{14.90}$$

et plus généralement, dans le cas non-réduit :

$$\boldsymbol{V}_{\hat{\theta}} = \sigma^2 (\boldsymbol{X}^t \boldsymbol{V}^{-1} \boldsymbol{X})^{-1} . \tag{14.91}$$

Notons que d'après (14.43) ces expressions peuvent également s'écrire :

$$\boldsymbol{V}_{\hat{\theta}} = 2\sigma^2 \boldsymbol{S}''^{-1} . \tag{14.92}$$

Variance des fonctions à estimer $\boldsymbol{\beta} = \boldsymbol{C\theta}$. Ceci englobe le cas régulier pour $\boldsymbol{C}$ quelconque et le cas singulier pour $\boldsymbol{CP} = \boldsymbol{0}$. On a, toujours d'après (14.88) :

$$\boldsymbol{V}_{C\hat{\theta}} = \sigma^2 \boldsymbol{C} (\boldsymbol{X}^t \boldsymbol{X})^{(-1)} \boldsymbol{C}^t . \tag{14.93}$$

14.3.10 Estimation de la variance σ^2.

Nous avons supposé que seule la matrice $\boldsymbol{V}$ des variances-covariances *relatives* de l'observation $\boldsymbol{y}$ était connue. Mais il réapparaît alors dans le calcul des variances-covariances des estimateurs $\widehat{\boldsymbol{\theta}}$ le facteur σ^2 fixant l'échelle absolue des variances-covariances de $\boldsymbol{y}$. La méthode des moindres carrés nous fournit un résidu $\widehat{\boldsymbol{\epsilon}}$, qui est une estimation du bruit $\boldsymbol{\epsilon}$; grâce à ce résidu, nous pouvons estimer σ^2. Considérons donc le résidu $\widehat{\boldsymbol{\epsilon}}$ de l'estimation de $\boldsymbol{\mu}$ par $\widehat{\boldsymbol{\mu}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$:

$$\widehat{\boldsymbol{\epsilon}} = \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\theta}} . \tag{14.94}$$

Nous avons vu que $\widehat{\boldsymbol{\epsilon}} = \boldsymbol{B}\boldsymbol{y} = \boldsymbol{B}\boldsymbol{\epsilon} = \widehat{\boldsymbol{\epsilon}}$. La somme des carrés des résidus, qui est la quantité $S_{\min}$ que nous avons minimisée, est égale à $S_{\min} = (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})^t (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) = \widehat{\boldsymbol{\epsilon}}^t \widehat{\boldsymbol{\epsilon}} = \boldsymbol{\epsilon}^t \boldsymbol{B}^t \boldsymbol{B} \boldsymbol{\epsilon} = \boldsymbol{\epsilon}^t \boldsymbol{B} \boldsymbol{\epsilon}$, ce qui peut encore s'écrire $\boldsymbol{\epsilon}^t \boldsymbol{B} \boldsymbol{\epsilon} = \mathrm{trace}(\boldsymbol{\epsilon}^t \boldsymbol{B} \boldsymbol{\epsilon})$. Comme la matrice $\boldsymbol{B}$ est symétrique ($\boldsymbol{B}^t = \boldsymbol{B}$), elle définit une forme quadratique, et on peut écrire : $S_{\min} = \mathrm{trace}(\boldsymbol{\epsilon}^t \boldsymbol{B} \boldsymbol{\epsilon}) = \mathrm{trace}(\boldsymbol{B} \boldsymbol{\epsilon} \boldsymbol{\epsilon}^t)$ soit, en prenant la valeur moyenne :

$$\begin{aligned} \mathrm{E}\{S_{\min}\} &= \mathrm{E}\{\mathrm{trace}(\boldsymbol{B}\boldsymbol{\epsilon}\boldsymbol{\epsilon}^t)\} = \mathrm{trace}(\boldsymbol{B}\,\mathrm{E}\{\boldsymbol{\epsilon}\boldsymbol{\epsilon}^t\}) \\ &= \sigma^2 \,\mathrm{trace}(\boldsymbol{B}) \\ &= \sigma^2 (n - r) , \end{aligned}$$

d'où finalement : $\mathrm{E}\{S_{\min}\} = \sigma^2 (n - r)$. On peut tirer de cette expression l'estimateur s^2 non-biaisé de σ^2 :

$$\begin{aligned} s^2 &= \frac{S_{\min}}{n - k} \quad \text{dans le cas régulier,} \\ s^2 &= \frac{S_{\min}}{n - r} \quad \text{dans le cas général,} \end{aligned} \tag{14.95}$$

où n est le nombre de valeurs observées, k le nombre de paramètres ajustables et r est le rang de la matrice du modèle ($\boldsymbol{X}$ ou $\boldsymbol{X}_\beta$).

Cette formule s'applique également dans le cas où les erreurs sont corrélées ($\mathrm{E}\{\boldsymbol{\epsilon}\boldsymbol{\epsilon}^t\} = \sigma^2\boldsymbol{V}$) et par conséquent $S_{\min} = (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\theta}})^t\boldsymbol{W}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\theta}})$, avec $\boldsymbol{W}^t = \boldsymbol{W} = \boldsymbol{V}^{-1}$.

▶**Exemple 14.7.** *Cas de l'ajustement d'une droite de moindres carrés.* Le modèle adopté est : $y_i = \theta_0 + \theta_1 x_i + \epsilon_i$, $\mathrm{E}\{\boldsymbol{\epsilon}\} = 0$ (pas d'erreurs systématiques), $\mathrm{E}\{\boldsymbol{\epsilon}\boldsymbol{\epsilon}^t\} = \boldsymbol{V} = \sigma^2\boldsymbol{I}$ (mesures d'égale précision). Dans ce cas la matrice modèle $\boldsymbol{X}$ vaut :

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \text{ il vient } \quad \boldsymbol{X}^t\boldsymbol{X} = \begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix}, \quad \boldsymbol{X}^t\boldsymbol{y} = \begin{pmatrix} \sum y_i \\ \sum x_i y_i \end{pmatrix},$$

$$(\boldsymbol{X}^t\boldsymbol{X})^{-1} = \frac{1}{n\sum x_i^2 - (\sum x_i)^2} \begin{pmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{pmatrix}.$$

D'où la solution $\widehat{\boldsymbol{\theta}} = (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{y}$:

$$\begin{pmatrix} \widehat{\theta}_0 \\ \widehat{\theta}_1 \end{pmatrix} = \frac{1}{n\sum x_i^2 - (\sum x_i)^2} \begin{pmatrix} \sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i \\ -\sum y_i \sum x_i + n \sum x_i y_i \end{pmatrix}, \tag{14.96}$$

soit, en posant : $\bar{x} = \frac{1}{n}\sum x_i$ et $\bar{y} = \frac{1}{n}\sum y_i$:

$$\widehat{\theta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \quad \text{et} \quad \bar{y} = \widehat{\theta}_0 + \widehat{\theta}_1 \bar{x}\ . \tag{14.97}$$

La dernière égalité montre que la droite des moindres carrés passe par le centre de gravité du nuage de points de coordonnées (x_i, y_i). Dans l'expression de la pente de cette droite, il faut noter la dissymétrie des rôles joués par les x_i et les y_i. Cela ne doit pas surprendre puisque les y_i sont des variables aléatoires alors que les x_i sont des valeurs sûres.

Calculons maintenant la matrice des variances-covariances de l'estimateur $\widehat{\boldsymbol{\theta}}$. Nous avons ici $\boldsymbol{V}_{\hat{\theta}} = \sigma^2(\boldsymbol{X}^t\boldsymbol{X})^{-1}$. L'expression $(\boldsymbol{X}^t\boldsymbol{X})^{-1}$ ayant déjà été trouvée quand on a calculé l'estimateur $\widehat{\boldsymbol{\theta}}$, le résultat est immédiat :

$$\boldsymbol{V}_{\hat{\theta}} = \frac{\sigma^2}{\sum (x_i - \bar{x})^2} \begin{pmatrix} \frac{1}{n}\sum x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix}. \tag{14.98}$$

En appliquant le résultat (14.95) précédent à l'ajustement du nuage de points (x_i, y_i) par la droite des moindres carrés on obtient l'estimation non-biaisée de σ^2 :

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n} (y_i - \widehat{\theta}_0 - \widehat{\theta}_1 x_i)^2 \tag{14.99}$$

14.3.11 Loi suivie par les estimateurs des moindres carrés.

Nous venons d'établir que, quelle que soit la taille n de l'échantillon, le vecteur aléatoire $\widehat{\boldsymbol{\theta}}$ est de moyenne $\boldsymbol{\theta}$ et de matrice des variances-covariances $\boldsymbol{V}_{\hat{\theta}}$. De même pour les fonctions à estimer, le vecteur aléatoire $\boldsymbol{C}\widehat{\boldsymbol{\theta}}$ est de moyenne $\boldsymbol{C}\boldsymbol{\theta}$ et de matrice des variances-covariances $\boldsymbol{V}_{C\hat{\theta}}$. Il reste à déterminer la loi asymptotique suivie par ces estimateurs.

Remarquons que $\widehat{\boldsymbol{\theta}}$ est obtenu comme combinaison linéaire des n variables réduites non-corrélées $\boldsymbol{y}$ et que par conséquent on doit s'attendre à ce que $\widehat{\boldsymbol{\theta}}$ tende assez vite vers la loi normale.

Cas régulier. Dans ce cas la matrice $\boldsymbol{X}^t\boldsymbol{X}$ est non-singulière et, si cette matrice reste non-singulière quand $n \to \infty$, on pourra appliquer la loi des grands nombres et montrer que $\widehat{\boldsymbol{\theta}}$ est un estimateur convergent de $\boldsymbol{\theta}$. Sous les mêmes hypothèses, le théorème central limite nous donne la loi asymptotique suivie par $\widehat{\boldsymbol{\theta}}$: il nous dit que $\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{\text{loi}} \mathcal{N}(\mathbf{0}, \sigma^2\boldsymbol{\Sigma}^{-1})$ où $\boldsymbol{\Sigma} = \lim_{n\to\infty} \frac{1}{n}\boldsymbol{X}^t\boldsymbol{X}$, si cette limite existe. En d'autres termes, pour n assez grand, la loi suivie par $\boldsymbol{\theta}$ est approximativement normale à k dimensions, de moyenne $\boldsymbol{\theta}$ et de matrice des variances-covariances $\sigma^2(\boldsymbol{X}^t\boldsymbol{X})^{-1}$. On ne peut pas étendre cette propriété pour $n \to \infty$ car les termes de $\boldsymbol{X}^t\boldsymbol{X}$ peuvent devenir infinis.

Pour les fonctions à estimer. Sous réserve que la matrice $\boldsymbol{V}_{C\hat{\theta}}$ reste non-singulière quand $n \to \infty$ on montrerait de même que $\boldsymbol{C}\widehat{\boldsymbol{\theta}}$ est un estimateur convergent de $\boldsymbol{C\theta}$ et que $\sqrt{n}(\boldsymbol{C}\widehat{\boldsymbol{\theta}}_n - \boldsymbol{C\theta}) \xrightarrow{\text{loi}} \mathcal{N}(\mathbf{0}, \sigma^2\boldsymbol{\Sigma}_C^{-1})$ où, si cette limite existe, $\boldsymbol{\Sigma}_C = \lim_{n\to\infty} \frac{1}{n}\boldsymbol{X}_\beta^t\boldsymbol{X}_\beta$.

Cas normal. Si l'observation $\boldsymbol{y}$ suit une loi normale, alors les mêmes résultats s'appliquent mais la loi suivie par les estimateurs $\widehat{\boldsymbol{\theta}}$ et $\boldsymbol{C}\widehat{\boldsymbol{\theta}}$ est normale pour n fini et non plus seulement asymptotiquement normale pour $n \to \infty$.

Ces résultats sont valables pour le cas réduit, pour le cas non-réduit (méthode des moindres carrés pondérés) il faudrait s'assurer que le changement de variables qui permet de se ramener au cas réduit soit toujours possible par passage à la limite infinie.

14.3.12 Région de confiance dans l'espace des paramètres.

La méthode des moindres carrés fournit un estimateur $\widehat{\boldsymbol{\theta}}$ du paramètre inconnu $\boldsymbol{\theta}$, cet estimateur dépend de l'échantillon et est, comme il se doit, une variable aléatoire. Il se pose alors le problème de l'estimation de la région où se répartit la variable aléatoire $\widehat{\boldsymbol{\theta}}$ lorsque l'échantillon $(y_1, \ldots, y_n)$ varie. Ce problème est celui de l'*estimation d'intervalle* qui a été traitée au chapitre 12 et plus particulièrement, dans le cas multidimensionnel, au paragraphe 12.4.

Par définition la zone de confiance $Q_\gamma(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})$ est une portion de l'espace des paramètres P_k, solution de l'équation :

$$\Pr\{\boldsymbol{\theta} \in Q_\gamma(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})\} = \gamma\,, \quad Q_\gamma \subset P_k\,, \tag{14.100}$$

où γ est une probabilité donnée à l'avance. En règle générale on choisit γ assez grand, par exemple $\gamma = 0.90$ et on parle d'intervalle à 90% de confiance. Il existe une infinité d'intervalles de confiance satisfaisant l'équation (14.100), parmi ceux-ci on choisit celui de plus petite surface c'est-à-dire celui dont la frontière est une courbe d'iso-densité de probabilité.

Pour résoudre entièrement le problème de la détermination de l'intervalle de confiance, il faut donc connaître la loi suivie par l'estimateur $\widehat{\boldsymbol{\theta}}$. Cette loi n'est connue que dans certains cas particuliers (dont le plus important est le cas linéaire et normal), en dehors de ces cas les résultats que nous allons établir ci-dessous ne seront qu'approximatifs.

Valeur de S au voisinage de $\widehat{\boldsymbol{\theta}}$.

L'estimateur des moindres carrés $\widehat{\boldsymbol{\theta}}$ est trouvé en minimisant la fonction $S(\boldsymbol{\theta})$. Développons $S(\boldsymbol{\theta})$ en série de Taylor au voisinage du minimum $\widehat{\boldsymbol{\theta}}$ pour lequel on a $S'(\widehat{\boldsymbol{\theta}}) = 0$. On obtient, de façon exacte dans le cas linéaire :

$$S(\boldsymbol{\theta}) = S(\widehat{\boldsymbol{\theta}}) + (\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})^t \boldsymbol{S}'(\widehat{\boldsymbol{\theta}}) + \frac{1}{2}(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})^t \boldsymbol{S}''(\widehat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}) ,$$

$$S(\boldsymbol{\theta}) = S(\widehat{\boldsymbol{\theta}}) + \frac{1}{2}(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})^t \boldsymbol{S}''(\widehat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}) ,$$

où $\boldsymbol{S}'$ est le vecteur gradient de S par rapport aux θ_i et $\boldsymbol{S}''$ le hessien de S (matrice des dérivées secondes de S). Il existe, dans le cas linéaire, une relation entre la matrice des variances-covariances $\boldsymbol{V}_{\hat{\theta}}$ de $\widehat{\boldsymbol{\theta}}$ et la valeur du hessien de S en $\widehat{\boldsymbol{\theta}}$. Cette relation est donnée par l'équation (14.92) : $\boldsymbol{V}_{\hat{\theta}} = 2\sigma^2 \boldsymbol{S}''(\widehat{\boldsymbol{\theta}})^{-1}$, d'où :

$$S(\boldsymbol{\theta}) = S(\widehat{\boldsymbol{\theta}}) + \sigma^2 (\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})^t \boldsymbol{V}_{\hat{\theta}}{}^{-1} (\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}) . \tag{14.101}$$

Posons $\Delta\boldsymbol{\theta} = \boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}$ égal à l'accroissement des paramètres autour de $\widehat{\boldsymbol{\theta}}$ et $\Delta S = S(\boldsymbol{\theta}) - S(\widehat{\boldsymbol{\theta}})$ égal à l'accroissement correspondant de la fonction S. Il vient :

$$\Delta S = \sigma^2 (\Delta\boldsymbol{\theta})^t \boldsymbol{V}_{\hat{\theta}}{}^{-1} \Delta\boldsymbol{\theta} . \tag{14.102}$$

La matrice $\boldsymbol{V}_{\hat{\theta}}$ étant définie positive l'équation (14.102) précédente est une forme quadratique définie positive.

Intervalle de confiance dans le cas linéaire et normal.

Dans le cas linéaire et normal, la variable aléatoire $\Delta\boldsymbol{\theta}$ est normale de moyenne nulle et de matrice des variances-covariances $\boldsymbol{V}_{\hat{\theta}}$. La quantité $\Delta S/\sigma^2$ suit alors une loi du χ^2 à k degrés de liberté (voir chapitre 6.3.9, page 101). La zone de confiance est alors entièrement déterminée par l'équation : $Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}) = S/\sigma^2 = k_\gamma^2$, où k_γ est une constante qui ne dépend que de la confiance γ et du nombre de paramètres k. On trouve k_γ^2 en inversant la fonction de répartition de la loi du χ^2 à k degrés de liberté :

$$k_\gamma^2 = F_{\chi^2}^{-1}(\gamma) . \tag{14.103}$$

La table 6.4 page 103, donne la valeur de k_γ pour certaines valeurs usuelles de γ et pour quelques valeurs du nombre de degrés de liberté qui ici est égal à k (le nombre de paramètres).

Interprétation géométrique de l'écart type σ_{mm} des estimateurs $\widehat{\theta}_m$.

Supposons que la matrice $\boldsymbol{V}_{\hat{\theta}}$ des variances-covariances de $\widehat{\boldsymbol{\theta}}$ est diagonale. Les éléments diagonaux de $\boldsymbol{V}_{\hat{\theta}}$ sont les variances des θ_i que nous noterons σ_{ii}^2. A partir du minimum $\widehat{\boldsymbol{\theta}}$, déplaçons-nous le long de l'axe θ_m, c'est-à-dire faisons varier θ_m en gardant les autres $\theta_{i\neq m}$ constants. Dans ce cas, $\Delta\boldsymbol{\theta}$ est un vecteur colonne partout nul sauf à la ligne $i = m$ où il vaut $\Delta\theta_m$, d'après (14.102) il vient :

$$\Delta S = \sigma^2 \frac{\Delta\theta_m^2}{\sigma_{mm}^2} \quad \text{ou} \quad \frac{\Delta S}{\sigma^2} = \frac{\Delta\theta_m^2}{\sigma_{mm}^2} . \tag{14.104}$$

La forme de cette équation suggère une méthode pratique pour calculer les écart types σ_{ii} des θ_i. Quand on a trouvé le point $\widehat{\boldsymbol{\theta}}$ où $S(\boldsymbol{\theta})$ est minimum, on forme la quantité $X^2(\boldsymbol{\theta}) = S(\boldsymbol{\theta})/\sigma^2$, et l'on se déplace à partir de $\widehat{\boldsymbol{\theta}}$ successivement le long de chaque axe θ_m d'une quantité $\Delta\theta_m$, jusqu'au point P où $X^2(\widehat{\boldsymbol{\theta}} + \Delta\boldsymbol{\theta}) = X^2(\widehat{\boldsymbol{\theta}}) + 1$. En ce point, d'après (14.104), la valeur absolue de $\Delta\theta_m$ est égale à l'écart type σ_{mm} de $\widehat{\theta}_m$.

Ce que nous avons montré, est que la forme quadratique définie positive : $X^2(\Delta\boldsymbol{\theta}) = 1$ atteint ses extrema, suivant chaque axe, pour la valeur σ_{ii}, c'est-à-dire :

$$\max_{\Delta\theta_i}\{\Delta\theta_i \,|\, (\Delta\boldsymbol{\theta})^t \boldsymbol{V}_{\hat{\theta}}{}^{-1} \Delta\boldsymbol{\theta} = 1\} = \sigma_{ii}\,, \tag{14.105a}$$

$$\min_{\Delta\theta_i}\{\Delta\theta_i \,|\, (\Delta\boldsymbol{\theta})^t \boldsymbol{V}_{\hat{\theta}}{}^{-1} \Delta\boldsymbol{\theta} = 1\} = -\sigma_{ii}\,, \tag{14.105b}$$

où σ_{ii} est le i^{e} élément diagonal de la matrice $\boldsymbol{V}_{\hat{\theta}}$. Cette propriété reste valable même si $\boldsymbol{V}_{\hat{\theta}}$ n'est pas diagonale mais est une matrice définie positive (théorème B.8, page 324). Les équations (14.105a) et (14.105b) ne dépendent donc pas des corrélations entre les variables θ_i (ce fait est illustré sur la figure 6.3 page 88).

Ces considérations conduisent finalement à l'interprétation géométrique suivante. Afin de trouver l'écart type σ_{mm} de l'estimateur $\widehat{\theta}_m$ dans le cas linéaire, il suffit de construire l'hyper-parallélépipède rectangle d'arêtes parallèles aux axes θ_i et circonscrit à l'hyper-ellipsoïde $X^2(\theta) = X^2_{\min} + 1$. Les arêtes de cet hyper-parallélépipède sont alors de longueur $2\sigma_{mm}$. Cette construction est illustrée par la figure 14.3, dans un cas à $k = 2$ dimensions.

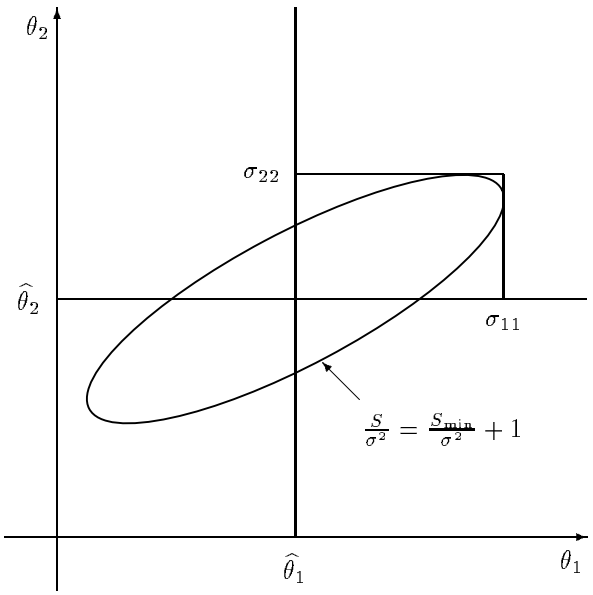


FIG. 14.3: *Construction géométrique des écart types σ_{11}, σ_{22} des estimateurs des moindres carrés $\widehat{\theta}_1$ et $\widehat{\theta}_2$, à partir de l'ellipse d'équation $S(\theta_1, \theta_2)/\sigma^2 = S_{\min}/\sigma^2 + 1$.*

Dans le cas linéaire et normal, la confiance γ qu'il faut accorder à la zone de confiance déterminée par l'hyper-ellipsoïde est donnée par : $\gamma = F_{\chi^2}(1)$. Comme le montre la table 14.1 cette confiance diminue rapidement avec le nombre de paramètres k (le nombre de degrés de liberté).

k	1	2	3	4	5	10
100γ	68.269	39.345	19.875	9.020	3.743	0.017

TAB. 14.1: *Confiance γ associée à l'hyper-ellipsoïde d'équation $S/\sigma^2 = S_{\min}/\sigma^2 + 1$, en fonction du nombre k de paramètres à ajuster.*

14.4 Résumé des propriétés du modèle linéaire.

Dans le modèle linéaire réduit et régulier où :

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}\,, \quad \mathrm{E}\{\boldsymbol{\epsilon}\} = \boldsymbol{0}\,, \quad \mathrm{E}\{\boldsymbol{\epsilon}\boldsymbol{\epsilon}^t\} = \sigma^2\boldsymbol{I}\,, \quad \det \boldsymbol{X}^t\boldsymbol{X} \neq 0\,, \tag{14.106}$$

les estimateurs des moindres carrés $\widehat{\boldsymbol{\theta}}$ sont linéaires et possèdent les propriétés suivantes :

1. Les estimateurs des moindres carrés sont convergents si la matrice $\frac{1}{n}\boldsymbol{X}^t\boldsymbol{X}$ est régulière lorsque la taille de l'échantillon augmente indéfiniment, c'est-à-dire :

$$[\lim_{n\to\infty} \det(\frac{1}{n}\boldsymbol{X}^t\boldsymbol{X}) \neq 0\,] \implies [\,\widehat{\boldsymbol{\theta}} \xrightarrow{\Pr} \boldsymbol{\theta}\,]\,. \tag{14.107}$$

   Le vecteur $\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ est alors asymptotiquement normal :

$$\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{\text{loi}} \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{\Sigma}^{-1})\,, \quad \text{où} \quad \boldsymbol{\Sigma} = \lim_{n\to\infty} \frac{1}{n}\boldsymbol{X}^t\boldsymbol{X}\,. \tag{14.108}$$

2. Les estimateurs $\widehat{\boldsymbol{\theta}}$ sont non-biaisés : $\mathrm{E}\{\widehat{\boldsymbol{\theta}}\} = \boldsymbol{\theta}$.

3. L'estimateur $s^2 = \frac{1}{n-k} S_{\min}$ de σ^2 est non- biaisé.

4. La matrice des variances-covariances de $\widehat{\boldsymbol{\theta}}$ est $\boldsymbol{V}_{\hat{\theta}} = \sigma^2(\boldsymbol{X}^t\boldsymbol{X})^{-1}$.

5. D'après le théorème de Gauss-Markov, l'estimateur $\widehat{\boldsymbol{\theta}}$ possède la plus petite variance dans la classe des estimateurs linéaires et sans biais de $\boldsymbol{\theta}$.

Si les erreurs $\boldsymbol{\epsilon}$ sont normales, l'estimateur $\widehat{\boldsymbol{\theta}}$ est un estimateur du maximum de vraisemblance et possède de plus les propriétés suivantes :

6. $\widehat{\boldsymbol{\theta}}$ possède la plus petite variance dans la classe des estimateurs sans biais de $\boldsymbol{\theta}$.

7. $\widehat{\boldsymbol{\theta}}$ suit un loi normale de moyenne $\boldsymbol{\theta}$ et de matrice des variances-covariances $\sigma^2(\boldsymbol{X}^t\boldsymbol{X})^{-1}$ c'est-à-dire : $(\widehat{\boldsymbol{\theta}}\text{-}\boldsymbol{\theta})$=$\mathcal{N}(\boldsymbol{0}, \sigma^2(\boldsymbol{X}^t\boldsymbol{X})^{-1})$.

8. La variable aléatoire $\frac{1}{\sigma^2}\widehat{\boldsymbol{\epsilon}}\widehat{\boldsymbol{\epsilon}}^t$, où $\widehat{\boldsymbol{\epsilon}} = \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\theta}}$, suit une loi du χ^2 à $n-k$ degrés de liberté.

9. Les estimateurs $\widehat{\boldsymbol{\theta}}$ et $s^2 = \frac{1}{n-k}\widehat{\boldsymbol{\epsilon}}\widehat{\boldsymbol{\epsilon}}^t = \frac{1}{n-k}S_{\min}$ sont exhaustifs pour $\boldsymbol{\theta}$ et σ^2.

10. $\widehat{\boldsymbol{\theta}}$ et s^2 sont indépendants.

11. La région de confiance autour de $\widehat{\boldsymbol{\theta}}$ où se trouve le vrai paramètre avec la probabilité γ est un hyper-ellipsoïde d'équation $\frac{1}{\sigma^2}(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})^t \boldsymbol{X}^t \boldsymbol{X}(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}) = k_\gamma^2$, où k_γ^2 est trouvé en inversant la fonction de répartition de la loi du χ^2 à k degrés de liberté : $k_\gamma^2 = F^{-1}(\gamma)$.

Les principaux résultats sont aussi indiqués dans les tables 14.2 et 14.3 où l'on envisage également le cas singulier :

- La table 14.2 donne la solution du modèle linéaire $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}$ par la méthode des moindres carrés.

- La table 14.3 donne la matrice des variances-covariances $\boldsymbol{V}_{\hat{\theta}}$ des estimateurs $\widehat{\boldsymbol{\theta}}$.

- Une estimation non-biaisée de σ^2 est donnée par :

$$\boxed{s^2 = \frac{S_{\min}}{n-r}, \quad r = \operatorname{rg} \boldsymbol{X}\,.} \tag{14.109}$$

  où n est le nombre de valeurs observées et r le rang de la matrice $\boldsymbol{X}$ du modèle.

	$\sigma^2 \boldsymbol{I}$	$\sigma^2 \boldsymbol{V}$
$\widehat{\boldsymbol{\theta}}$	$(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{y}$	$(\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{y}$
$\boldsymbol{C}\widehat{\boldsymbol{\theta}}$	$\boldsymbol{C}(\boldsymbol{X}^t\boldsymbol{X})^{(-1)}\boldsymbol{X}^t\boldsymbol{y}$	$\boldsymbol{C}(\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X})^{(-1)}\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{y}$

TAB. 14.2: *Solutions du modèle linéaire par la méthode des moindres carrés. La notation $^{(-1)}$ désigne la matrice pseudo-inverse et les fonctions à estimer $\boldsymbol{\beta} = \boldsymbol{C}\boldsymbol{\theta}$ ne sont définies que si $\boldsymbol{C}\boldsymbol{P} = \boldsymbol{0}$, où $\boldsymbol{P}$ est un projecteur orthogonal sur le noyau de $\boldsymbol{X}^t\boldsymbol{X}$ (ou de manière équivalente sur le noyau de $\boldsymbol{X}$). Une condition nécessaire et suffisante pour que la pseudo-inverse soit égale à l'inverse classique est que les colonnes de $\boldsymbol{X}$ soient linéairement indépendantes : dans ce cas $\boldsymbol{C}$ peut être quelconque.*

	$\sigma^2 \boldsymbol{I}$	$\sigma^2 \boldsymbol{V}$
$\boldsymbol{V}_{\hat{\theta}}$	$\sigma^2(\boldsymbol{X}^t\boldsymbol{X})^{-1}$	$\sigma^2(\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}$
$\boldsymbol{V}_{C\hat{\theta}}$	$\sigma^2\boldsymbol{C}(\boldsymbol{X}^t\boldsymbol{X})^{(-1)}\boldsymbol{C}^t$	$\sigma^2\boldsymbol{C}(\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X})^{(-1)}\boldsymbol{C}^t$

TAB. 14.3: *Matrice des variances-covariances des paramètres estimés par la méthode des moindres carrés. L'introduction de matrices pseudo-inverses appelle les mêmes remarques que pour la table 14.2*

14.5 Exercices et problèmes.

Exercice 14.1. Trouver les pseudo-solutions normales des systèmes singuliers suivants : 1) $0x = 0$; 2) $x = a$, $x = b$; 3) $x + y = a$.

Exercice 14.2. Montrer que les estimateurs $\widehat{\boldsymbol{\beta}} = \boldsymbol{X}^{(-1)}\boldsymbol{X}\widehat{\boldsymbol{\theta}}$ sont non-biaisés pour l'estimation de $\boldsymbol{\beta} = \boldsymbol{X}^{(-1)}\boldsymbol{X}\boldsymbol{\theta}$.

Exercice 14.3. Démontrer que la somme des carrés des éléments des lignes (ou des colonnes) de la matrice $\boldsymbol{H}$ est égale à l'élément diagonal correspondant, c'est-à-dire $\sum_{j=1}^{n} h_{ij}^2 = h_{ii}$.

Exercice 14.4. Sans faire appel au théorème de Gauss-Markov, démontrer que les estimateurs $\boldsymbol{C}\widehat{\boldsymbol{\theta}}$ de $\boldsymbol{C}\boldsymbol{\theta}$ sont non-biaisés. Toujours sans faire appel à ce théorème, calculer l'expression de la matrice des variances-covariances de ces estimateurs.

Exercice 14.5. *Ajustement par une fonction quelconque.* Généraliser les résultats obtenus dans l'exemple 14.7 aux modèles du type $y_i = \theta_0 + \theta_1 f(x_i) + \epsilon_i$, (par exemple $y_i = \theta_0 + \theta_1 \ln(x_i) + \epsilon_i$). Montrer que l'estimation des θ_i est donnée par :

$$\widehat{\theta}_1 = \frac{\sum(x_2 - \bar{x}_2)(y_2 - \bar{y})}{\sum(x_2 - \bar{x}_2)^2}\,, \quad \bar{y} = \widehat{\theta}_0 + \widehat{\theta}_1\bar{x}_2\,.$$

Dans ces formules les indices de sommation sont omis, et les expressions du type $\sum x_2$ représentent la somme de tous les éléments de la deuxième colonne de la matrice $\boldsymbol{X}$, tandis que $\bar{x}_2$ représente la moyenne des éléments de cette deuxième colonne. Avec ces conventions, trouver que la matrice des variances-covariances des estimateurs est donnée par :

$$\boldsymbol{V}_{\hat{\theta}} = \frac{\sigma^2}{\sum(x_2 - \bar{x}_2)^2}\begin{pmatrix} \frac{1}{n}\sum x_2^2 & -\bar{x}_2 \\ -\bar{x}_2 & 1 \end{pmatrix}$$

Chapitre 15

Estimation des paramètres de certaines lois.

15.1 Une loi à un paramètre : la loi exponentielle.

Cherchons, à l'aide d'une réalisation $(x_1, \ldots, x_n)$ d'un n-échantillon formé de valeurs indépendantes de la variable aléatoire X, à estimer le paramètre θ de la loi parente supposée être exponentielle :

$$f(x;\theta) = \frac{1}{\theta} e^{-x/\theta} \ . \tag{15.1}$$

C'est une loi de moyenne et de variance :

$$\mathrm{E}\{X\} = \theta, \quad \mathrm{Var}(X) = \theta^2 \, . \tag{15.2}$$

15.1.1 Estimation ponctuelle.

Calculons la fonction de vraisemblance :

$$L(x_1, \ldots, x_n|\theta) = \frac{1}{\theta^n} \exp\Big(-\frac{1}{\theta} \sum_{i=1}^{n} x_i\Big) \ . \tag{15.3}$$

La statistique $\sum X_i$ est exhaustive, puisque la fonction de vraisemblance du n-échantillon ne dépend des X_i que par l'intermédiaire de cette statistique. Calculons $\partial \ln L / \partial \theta$:

$$\ln L = -n \ln \theta - \frac{1}{\theta} \sum X_i \ ,$$

$$\frac{\partial \ln L}{\partial \theta} = -\frac{n}{\theta} + \frac{1}{\theta^2} \sum X_i = \frac{n}{\theta^2} \Big(\frac{1}{n} \sum X_i - \theta\Big) \ .$$

Ce résultat montre de plus que la statistique non-biaisée $\frac{1}{n} \sum X_i = \bar{X}$ est efficace MVB pour l'estimation de θ. C'est également, comme on devait s'y attendre, la solution donnée par la méthode du maximum de vraisemblance :

$$\left.\frac{\partial \ln L}{\partial \theta}\right|_{\theta=\bar{x}} = 0 \quad \text{et,} \quad \left.\frac{\partial^2 \ln L}{\partial \theta^2}\right|_{\theta=\bar{x}} = -\frac{n}{\bar{x}^2} < 0 \tag{15.4}$$

L'information de Fisher contenue dans le n-échantillon est égale à :

$$I_n(\theta) = \frac{n}{\theta^2}\,. \tag{15.5}$$

La moyenne de l'échantillon étant toujours non-biaisée, quand la moyenne de la population existe, on a alors :

$$\mathrm{E}\left\{\bar{X}\right\} = \theta\,. \tag{15.6}$$

On obtient la variance de $\bar{X}$, dans le cas MVB, par l'inverse de l'information de Fisher :

$$\mathrm{Var}(\bar{X}) = \frac{\theta^2}{n}\,. \tag{15.7}$$

Nous prendrons donc comme estimation ponctuelle de θ :

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i\,. \tag{15.8}$$

15.1.2 Estimation d'intervalle.

Montrons que la variable aléatoire $Y = 2\sum_{i=1}^{n} X_i/\theta$ suit une loi du χ^2 à $2n$ degrés de liberté. On sait par ailleurs, équation (9.11), que la variable aléatoire $t = \sum_{i=1}^{n} X_i$ suit une loi gamma, de densité de probabilité :

$$f(t;\theta) = \frac{\theta^{-n}}{\Gamma(n)} t^{n-1} e^{-t/\theta}\, H(t) \quad \text{où } H \text{ est la fonction de Heaviside.} \tag{15.9}$$

Effectuons le changement de variable $Y = 2t/\theta$. La densité de probabilité $g(x)$ de Y, est telle que $f(t)dt = g(x)dx$. Il vient :

$$\begin{aligned}
\frac{dt}{dx} &= \frac{\theta}{2} \quad \text{et donc} \\
g(x) &= \frac{\theta}{2}\frac{\theta^{-n}}{\Gamma(n)} t^{n-1} e^{-t/\theta}\, H(t); \quad \text{mais } H(t) = H(x) \\
&= \frac{1}{2\Gamma(n)} \left(\frac{t}{\theta}\right)^{n-1} e^{-t/\theta}\, H(x); \quad \text{et comme } \frac{t}{\theta} = \frac{x}{2} \\
g(x) &= \frac{1}{2^n \Gamma(n)} x^{n-1} e^{-x/2}\, H(x)\,.
\end{aligned}$$

On reconnaît là une loi du χ^2 à $2n$ degrés de liberté. L'intervalle bilatéral symétrique de la variable aléatoire Y au niveau γ est donné par l'équation :

$$\Pr\left\{\chi^2_{\frac{1}{2}+\frac{\gamma}{2}} \le Y < \chi^2_{\frac{1}{2}-\frac{\gamma}{2}}\right\} = \gamma\,, \tag{15.10}$$

$$\text{mais,} \quad x = \frac{2}{\theta}t, \quad \text{et} \quad \hat{\theta} = \frac{1}{n}t\,, \quad \text{soit:} \quad x = 2n\frac{\hat{\theta}}{\theta}\,.$$

L'équation devient :

$$\Pr\left\{\chi^2_{\frac{1}{2}+\frac{\gamma}{2}} \le 2n\frac{\hat{\theta}}{\theta} < \chi^2_{\frac{1}{2}-\frac{\gamma}{2}}\right\} = \gamma\,,$$

d'où l'on déduit l'intervalle de confiance sur θ calculé à partir de l'estimation $\hat{\theta}$:

$$\boxed{\Pr\left\{\hat{\theta}\frac{2n}{\chi^2_{\frac{1}{2}-\frac{\gamma}{2}}} < \theta \leq \hat{\theta}\frac{2n}{\chi^2_{\frac{1}{2}+\frac{\gamma}{2}}}\right\} = \gamma\,.} \tag{15.11}$$

Les valeurs telles que $\chi^2_{\frac{1}{2}-\frac{\gamma}{2}}$ sont les quantiles de la loi du χ^2 à $2n$ degrés de liberté. On rappelle que les quantiles x_α d'une loi de densité de probabilité $f(x)$ sont définis par l'équation :

$$\alpha = \int_{x_\alpha}^{\infty} f(u)\,du\,. \tag{15.12}$$

15.2 Une loi à deux paramètres : la loi normale.

Soit une réalisation $(x_1, \ldots, x_n)$ de n valeurs indépendantes de la variable aléatoire X, suivant la loi normale de densité de probabilité :

$$f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}\exp -\frac{(x-\mu)^2}{2\sigma^2}\,\cdot \tag{15.13}$$

C'est une loi à deux paramètres μ et σ. Il y a donc lieu d'envisager 5 cas possibles.

15.2.1 Estimation de la moyenne μ connaissant σ.

La statistique $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ est un estimateur convergent, non-biaisé et efficace (MVB) du paramètre μ. On a :

$$\mathrm{E}\left\{\bar{X}\right\} = \mu, \quad \mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}\,\cdot \tag{15.14}$$

La variable aléatoire :

$$y = \frac{\bar{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}} = \sqrt{n}\frac{\bar{X}-\mu}{\sigma}\,, \tag{15.15}$$

est une variable aléatoire normale réduite $\mathcal{N}(0,1)$. On en déduit, la probabilité γ étant donnée, l'intervalle de variation bilatéral symétrique de cette variable aléatoire :

$$\Pr\left\{y_{\frac{1}{2}+\frac{\gamma}{2}} < \sqrt{n}\frac{\bar{X}-\mu}{\sigma} \leq y_{\frac{1}{2}-\frac{\gamma}{2}}\right\} = \gamma\,. \tag{15.16}$$

Les quantités telles que $y_{\frac{1}{2}+\frac{\gamma}{2}}$ sont trouvées dans une table des quantiles de la loi normale réduite, d'où l'intervalle de confiance sur μ :

$$\boxed{\Pr\left\{\bar{X} - \frac{\sigma}{\sqrt{n}}y_{\frac{1}{2}-\frac{\gamma}{2}} \leq \mu < \bar{X} - \frac{\sigma}{\sqrt{n}}y_{\frac{1}{2}+\frac{\gamma}{2}}\right\} = \gamma\,.} \tag{15.17}$$

Comme la loi normale est paire, $y_{\frac{1}{2}-\frac{\gamma}{2}} = -y_{\frac{1}{2}+\frac{\gamma}{2}}$. On obtient alors pour une observation $\bar{x}$ de $\bar{X}$:

$$\mu = \bar{x} \pm \frac{\sigma}{\sqrt{n}}y_{\frac{1}{2}-\frac{\gamma}{2}}, \quad \text{avec la confiance } \gamma\,. \tag{15.18}$$

Application numérique Supposons que nous disposions d'un échantillon de taille 10 issu d'une loi normale :

$$x_1 = 0.621 \quad x_2 = 0.544 \quad x_3 = 1.252 \quad x_4 = -1.470 \quad x_5 = -1.131$$
$$x_6 = -0.830 \quad x_7 = 2.036 \quad x_8 = -0.135 \quad x_9 = 2.041 \quad x_{10} = 0.840,$$

et que nous sachions par ailleurs que $\sigma = 1$. On calcule $\bar{x} = 0.3768$ et on obtient, pour deux valeurs de γ, le tableau suivant :

γ	0.98	0.5
$y_{\frac{1}{2}-\frac{\gamma}{2}}$	$y_{0.01}$ =2.32635	$y_{0.25}$ =0.67449
μ_{inf}	-0.359	0.164
μ_{sup}	1.113	0.590

15.2.2 Estimation de μ ne connaissant pas σ.

La statistique $\bar{X}$ est encore un estimateur convergent, sans biais et efficace MVB de μ. On a :

$$\mathrm{E}\left\{\bar{X}\right\} = \mu, \quad \mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}\,. \tag{15.19}$$

Comme maintenant nous ne connaissons pas σ^2, il faut donc l'estimer. La statistique :

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{X})^2, \tag{15.20}$$

est un estimateur convergent, sans biais et asymptotiquement efficace de σ^2. On a :

$$\mathrm{E}\left\{S^2\right\} = \sigma^2, \quad \mathrm{Var}(S^2) = \frac{2\sigma^4}{n-1}\ \cdot \tag{15.21}$$

La variable aléatoire :

$$y = \frac{\bar{X} - \mu}{\sqrt{\dfrac{S^2}{n}}} = \sqrt{n}\frac{\bar{X} - \mu}{S}\,,$$

suit une loi de Student à $(n-1)$ degrés de liberté, d'où l'on déduit l'intervalle de confiance sur μ :

$$\boxed{\Pr\left\{\bar{X} - \frac{S}{\sqrt{n}}t_{\frac{1}{2}-\frac{\gamma}{2}} \le \mu < \bar{X} - \frac{S}{\sqrt{n}}t_{\frac{1}{2}+\frac{\gamma}{2}}\right\} = \gamma} \tag{15.22}$$

Les valeurs telles que $t_{\frac{1}{2}-\frac{\gamma}{2}}$ sont les quantiles de la loi de Student à $(n-1)$ degrés de liberté et, puisque la loi de Student est paire : $t_{\frac{1}{2}+\frac{\gamma}{2}} = -t_{\frac{1}{2}-\frac{\gamma}{2}}$. Il en résulte, à partir d'une observation $\bar{x}$ et s^2 :

$$\mu = \bar{x} \pm \frac{s}{\sqrt{n}}t_{\frac{1}{2}-\frac{\gamma}{2}}, \quad \text{avec la confiance } \gamma \tag{15.23}$$

Application numérique. On calcule avec les valeurs précédentes $\bar{x} = 0.3768$, $s^2 = 1.5548$, $s = 1.2469$ et l'on obtient le tableau suivant :

γ	0.98	0.5
$t_{\frac{1}{2}-\frac{\gamma}{2}}(f=9)$	$t_{0.01}$ =2.821	$t_{0.25}$ =0.703
μ_{inf}	-0.736	0.100
μ_{sup}	1.489	0.654

15.2.3 Estimation de σ^2 connaissant μ.

La statistique :

$$S'^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 \tag{15.24}$$

est un estimateur convergent, sans biais et efficace MVB de σ^2. On a :

$$\mathrm{E}\left\{S'^2\right\} = \sigma^2, \quad \mathrm{Var}(S'^2) = \frac{2\sigma^4}{n}. \tag{15.25}$$

La variable aléatoire :

$$\chi^2 = n\frac{S'^2}{\sigma^2}, \tag{15.26}$$

suit une loi du χ^2 à n degrés de liberté. On en déduit l'intervalle de confiance sur σ^2 :

$$\boxed{\Pr\left\{\frac{nS'^2}{\chi^2_{\frac{1}{2}-\frac{\gamma}{2}}} \leq \sigma^2 < \frac{nS'^2}{\chi^2_{\frac{1}{2}+\frac{\gamma}{2}}}\right\} = \gamma\,.} \tag{15.27}$$

Les valeurs telles que $\chi^2_{\frac{1}{2}-\frac{\gamma}{2}}$ sont les quantiles de la loi du χ^2 à n degrés de liberté.

Application numérique. On sait par ailleurs que $\mu = 0$. On calcule alors avec les valeurs précédentes $s'^2 = 1.54128$, $ns'^2 = 15.4128$ et l'on obtient le tableau suivant :

γ	0.98	0.5
$\chi^2_{\frac{1}{2}-\frac{\gamma}{2}}(f=10)$	$\chi^2_{0.01}$ =23.209	$\chi^2_{0.25}$ =12.549
$\chi^2_{\frac{1}{2}+\frac{\gamma}{2}}(f=10)$	$\chi^2_{0.99}$ =2.558	$\chi^2_{0,75}$ =6.737
σ^2_{inf}	0.664	1.228
σ^2_{sup}	6.025	2.288

15.2.4 Estimation de σ^2 ne connaissant pas μ.

Le paramètre μ étant inconnu, il faut l'estimer. La statistique $\bar{X}$ est encore un estimateur convergent, sans biais et efficace MVB de μ. La statistique :

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2\,, \tag{15.28}$$

est un estimateur convergent, sans biais et asymptotiquement efficace de σ^2. On a :

$$\mathrm{E}\left\{S^2\right\} = \sigma^2, \quad \mathrm{Var}(S^2) = \frac{2\sigma^4}{n-1}\ \cdot \tag{15.29}$$

La variable aléatoire :

$$\chi^2 = (n-1)\frac{s^2}{\sigma^2}\,, \tag{15.30}$$

suit une loi du χ^2 à $(n-1)$ degrés de liberté, d'où l'on déduit l'intervalle de confiance sur σ^2 :

$$\boxed{\Pr\left\{\frac{(n-1)S^2}{\chi^2_{\frac{1}{2}-\frac{\gamma}{2}}} \leq \sigma^2 < \frac{(n-1)S^2}{\chi^2_{\frac{1}{2}+\frac{\gamma}{2}}}\right\} = \gamma\,.} \tag{15.31}$$

Les valeurs telles que $\chi^2_{\frac{1}{2}-\frac{\gamma}{2}}$ sont les quantiles de la loi du χ^2 à $(n-1)$ degrés de liberté.

Application numérique. On calcule avec les valeurs précédentes $\bar{x} = 0.3768$, $s^2 = 1.5548$, $(n-1)s^2 = 13.9931$ et l'on obtient le tableau suivant :

γ	0.98	0.5
$\chi^2_{\frac{1}{2}-\frac{\gamma}{2}}(f=9)$	$\chi^2_{0.01}$ =21.666	$\chi^2_{0.25}$ =11.389
$\chi^2_{\frac{1}{2}+\frac{\gamma}{2}}(f=9)$	$\chi^2_{0.99}$ =2.088	$\chi^2_{0.75}$ =6.899
σ^2_{inf}	0.646	1.229
σ^2_{sup}	6.702	2.028

15.2.5 Estimation simultanée de μ et σ^2.

Le couple aléatoire $(\bar{X}, S^2)$, défini par :

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i\ , \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2, \tag{15.32}$$

est dans ce cas formé de variables aléatoires indépendantes. C'est un estimateur convergent, sans biais, asymptotiquement efficace, du couple (μ, σ^2) ; il possède

un vecteur moyenne μ et une matrice des variances-covariances $\boldsymbol{V}$, donnés par les expressions :

$$\boldsymbol{\mu} = \begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix}, \qquad \boldsymbol{V} = \begin{pmatrix} \dfrac{\sigma^2}{n} & 0 \\ 0 & \dfrac{2\sigma^4}{n-1} \end{pmatrix}. \tag{15.33}$$

La densité de probabilité du couple : $(\bar{X}, (n-1)S^2/\sigma^2)$ est donc le produit d'une loi normale $N(\mu, \sigma^2/n)$ par une loi du χ^2 à $(n-1)$ degrés de liberté. Dans le cas où n est petit, il est délicat de trouver analytiquement la région de confiance dans le plan μ, σ^2 ; mais pour n assez grand, cette densité de probabilité tend rapidement vers le produit de deux lois normales :

$$N(\mu, \sigma^2/n) \otimes N(n-1, 2(n-1)), \tag{15.34}$$

ce qui donne en se ramenant à $(\bar{X}, S^2)$ par changement de variables :

$$N(\mu, \sigma^2/n) \otimes N(\sigma^2, 2\sigma^4/(n-1)) . \tag{15.35}$$

C'est une loi normale à deux dimensions de forme quadratique associée :

$$Q(\bar{x}, s^2; \mu, \sigma^2) = \frac{n(\bar{x}-\mu)^2}{\sigma^2} + \frac{(n-1)(s^2-\sigma^2)^2}{2\sigma^4} \quad . \tag{15.36}$$

Les courbes telles que $Q = \lambda^2$ sont des ellipses contenant la probabilité :

$$P(\lambda) = 1 - \exp\left(-\lambda^2/2\right) . \tag{15.37}$$

Au niveau de confiance γ, la région du plan $\bar{x}, s^2$ délimitée par l'équation $Q(\bar{x}, s^2|\mu, \sigma^2) \leq \lambda^2$ est le domaine de variation du couple aléatoire $(\bar{x}, s^2)$, alors que la région du plan μ, σ^2 délimitée par l'équation $Q(\mu, \sigma^2|\bar{x}, s^2) \leq \lambda^2$ est la région de confiance du point (μ, σ^2). Pour que la région de confiance soit fermée, il faut remplir la condition :

$$(n-1) > 2\lambda^2 \equiv 4\ln(1-\gamma)^{-1} . \tag{15.38}$$

Application numérique. On calcule, toujours avec les 10 valeurs précédentes $\bar{x} = 0.3768$, $s^2 = 1.5548$, $(n-1)s^2 = 13.9931$. Au niveau $\gamma = 0.98$, $\lambda^2 = 7.82$ et la courbe n'est pas fermée ; en revanche au niveau $\gamma = 0.5$, $\lambda^2 = 1.386$, la courbe est fermée. Cette courbe est donnée par la figure 15.1. Notons qu'il faut un échantillon de taille supérieure ou égale à 17 pour que la région de confiance soit fermée au niveau de confiance $\gamma = 0.98$. Les courbes délimitant les régions de confiance sont loin d'être des ellipses, ce qui est normal pour un échantillon de si petite taille.

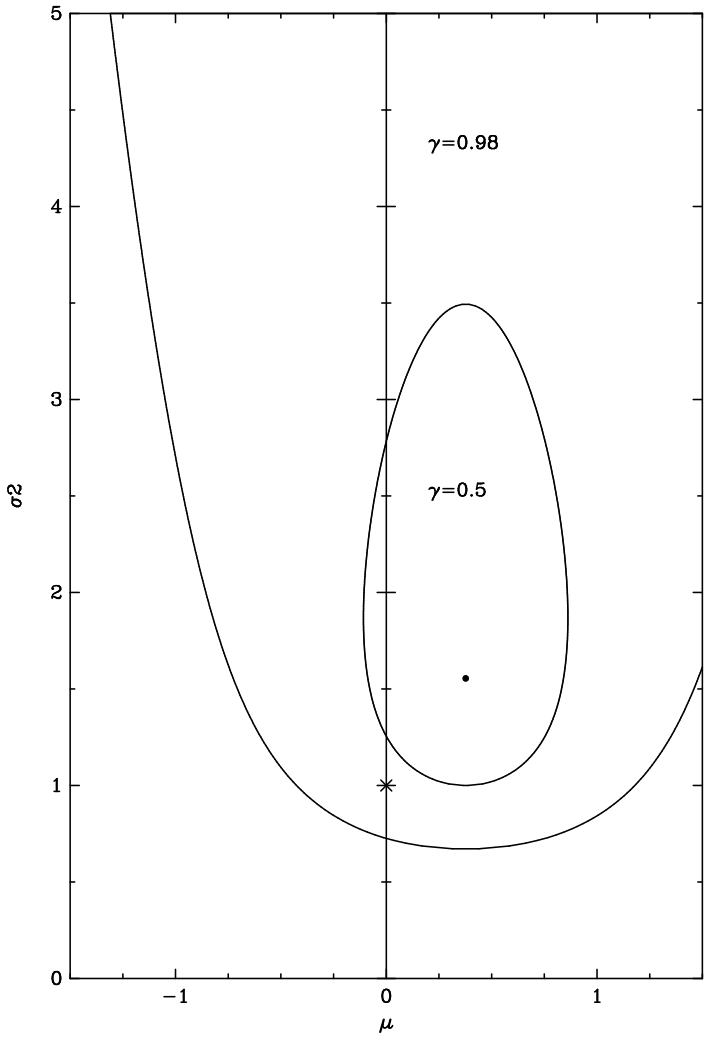


FIG. 15.1: *Région de confiance au niveau $\gamma = 0.5$ de l'estimation simultanée de la moyenne et de la variance d'une population normale, calculée à partir d'un échantillon de taille 10, Le point $\bullet$ indique la position de l'estimateur ponctuel $(\bar{x}, s^2)$, et la croix $\times$ indique les vraies valeurs du couple (μ, σ^2).*

Chapitre 16

Estimation de la loi.

Nous nous intéressons dans ce chapitre à l'estimation de la loi suivie par les variables aléatoires X_i d'un échantillon i.i.d $(X_1, \ldots, X_n)$, lorsque l'on dispose d'une réalisation $(x_1, \ldots, x_n)$ de cet échantillon.

16.1 Estimation de la fonction de répartition.

Soit $F(x)$ la fonction de répartition inconnue de la population parente d'où est issu $(X_1, \ldots, X_n)$.

16.1.1 L'estimateur « naturel » F_n.

Nous avons défini au chapitre 9.4 la fonction de répartition empirique F_n, calculée à partir d'un échantillon i.i.d $(X_1, \ldots, X_n)$, comme étant pour tout x le nombre de variables X_i n'ayant pas dépassé le seuil x. Soit :

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{]-\infty, x]}(X_i) \,. \tag{16.1}$$

Rappelons brièvement les résultats exposés dans ce chapitre.

1. La variable aléatoire $F_n(x)$ est une variable aléatoire discrètre à valeurs dans $\{0, \frac{1}{n}, \frac{2}{n}, \ldots, 1\}$.

2. La variable aléatoire $nF_n(x)$ suit une loi binomiale de paramètre $p = F(x)$:

$$nF_n(x) = \mathcal{B}(n, F(x)) \,. \tag{16.2}$$

3. Les variables aléatoires $nF_n(x)$ et $F_n(x)$ possèdent des moments à tous les ordres. En particulier $F_n(x)$ possède pour tout x une moyenne et une variance :

$$\mathrm{E}\{F_n(x)\} = F(x) \,, \quad \mathrm{Var}(F_n(x)) = \frac{1}{n} F(x)(1 - F(x)) \,. \tag{16.3}$$

L'estimateur $F_n(x)$ est donc non-biaisé pour l'estimation de $F(x)$.

4. La variable aléatoire $F_n(x)$ converge presque-sûrement vers $F(x)$ lorsque $n \to \infty$ et cette convergence est uniforme en x (théorème 9.1 de Glivenko-Cantelli) :

$$\sup_x |F_n(x) - F(x)| \xrightarrow{\text{p.s.}} 0 \,. \tag{16.4}$$

   L'estimateur $F_n(x)$ est donc convergent pour $F(x)$. Il suffisait d'ailleurs que $F_n(x)$ converge seulement *en probabilité* vers $F(x)$ pour être convergent.

5. La variable aléatoire $F_n(x)$ converge en loi vers une loi normale :

$$F_n(x) \xrightarrow{\text{loi}} \mathcal{N}(F(x), \tfrac{1}{n}F(x)(1-F(x))) \,. \tag{16.5}$$

   L'estimateur $F_n(x)$ est donc asymptotiquement efficace.

Bien que l'estimateur $F_n(x)$ présente toutes les caractéristiques d'un *bon* estimateur, la dernière propriété est asymptotique et nous ne savons pas comment, pour n fini, $F_n(x)$ converge vers $F(x)$. L'équation (16.5) nous dit que, pour n assez grand, l'erreur entre F_n et F peut être rendue aussi petite que l'on veut avec une probabilité elle aussi arbitrairement petite ; c'est-à-dire :

$$\forall \epsilon, \delta > 0,\ \exists N > 0,\ \text{ tel que } \Pr\Big\{ \max_{n \geq N} \sup_x |F_n(x) - F(x)| \leq \epsilon \Big\} \geq 1 - \delta \,, \tag{16.6}$$

mais elle ne nous renseigne pas sur la façon dont N dépend de ϵ et δ. En particulier nous ne savons pas quelle taille N doit avoir l'échantillon afin que l'on soit sûr, avec une probabilité δ de se tromper, que l'écart maximum entre F_n et F ne soit pas supérieur à ϵ. Ce problème a été abordé par A. N. Kolmogorov et fait l'objet du chapitre suivant.

16.1.2 La statistique de Kolmogorov.

La statistique $\sup_x |F_n(x) - F(x)|$ introduite ci-dessus mesure la « distance » entre les fonctions empirique F_n et théorique F. Cette distance découle de la norme de la convergence uniforme bien connue en analyse. En tant que variable aléatoire nous nommerons cette quantité « statistique de Kolmogorov » et nous la noterons D_n :

$$D_n = \sup_x |F_n(x) - F(x)| \,. \tag{16.7}$$

Cette statistique a été utilisée par Smirnov afin de juger de l'adéquation de F_n avec F dans ce que l'on appelle le « test de Kolmogorov-Smirnov ». Pour trouver D_n, il n'est pas nécessaire de chercher le maximum sur tous les x, il suffit de le chercher aux valeurs de l'échantillon. En effet :

$$\begin{aligned} D_n &= \max_{1 \leq i \leq n} (|F_n(X^+_{(i)}) - F(X_{(i)})|, |F_n(X^-_{(i)}) - F(X_{(i)})|) \\ &= \max_{1 \leq i \leq n} (|\frac{i}{n} - F(X_{(i)})|, |\frac{i-1}{n} - F(X_{(i)})|) \,, \end{aligned}$$

ou encore, en considérant le point milieu :

$$D_n = \max_{1 \leq i \leq n} (|\frac{2i-1}{2n} - F(X_{(i)})|) + \frac{1}{2n} \,. \tag{16.8}$$

Kolmogorov a montré que la variable aléatoire $Z = \sqrt{n} D_n$ suivait une loi indépendante de F quand $n \to \infty$. Plus précisément, on dispose du théorème suivant :

Théorème de Kolmogorov. *Si F est une fonction de répartition continue, alors :*

$$\lim_{n\to\infty} \Pr\left\{\sqrt{n}\sup_x |F_n(x) - F(x)| \le z\right\} = K(z), \tag{16.9}$$

où $K(z)$ est une fonction indépendante de F appelée fonction de répartition de Kolmogorov. Elle a pour expression :

$$K(z) = \sum_{k=-\infty}^{+\infty} (-1)^k e^{-2k^2z^2} \ . \tag{16.10}$$

On a aussi pour n fini :

$$\Pr\left\{\sqrt{n}D_n \le z\right\} = K(z)(1 - \frac{2k^2z}{3\sqrt{n}} + o(\frac{1}{n})) \ . \tag{16.11}$$

Le graphe de la fonction $K(z)$ est représenté sur la figure 16.1.

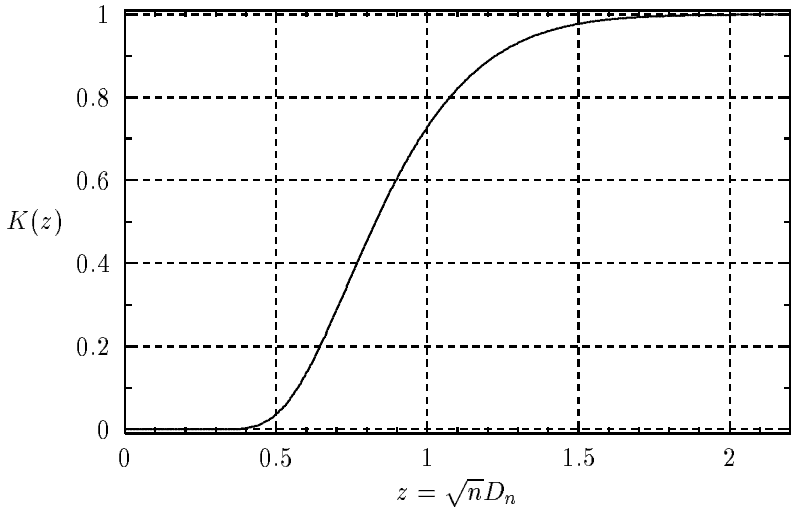


FIG. 16.1: *Fonction de répartition de Kolmogorov $K(z)$. La statistique D_n est la statistique de Kolmogorov et n est la taille de l'échantillon.*

16.2 Estimation d'une loi en présence de données censurées.

Il arrive souvent, dans la pratique, que l'échantillon $(X_1, \ldots, X_n)$ à partir duquel on tente d'estimer la loi F, soit composé de données hétérogènes : 1) des valeurs qui sont de véritables observations de la variable aléatoire X ; 2) des valeurs qui ne sont que des bornes supérieures (ou inférieures). On peut être tenté de ne garder que les observations au sens strict et d'estimer F_n par (16.1) par exemple. Cette façon de faire présente l'inconvénient de réduire la taille de l'échantillon et on se prive par ailleurs d'une information qui est en fait exploitable.

16.2.1 Modèle de censure.

Supposons que l'on cherche la loi suivie par le temps Y au bout duquel apparaît un certain événement B après un certain autre A. Le phénomène A, qui fixe l'origine des temps, peut être l'explosion d'une super-nova et B l'apparition d'un pulsar. On a observé un échantillon de n étoiles ou restes de super-novæ, et pour chaque observation i, $i = 1, \ldots, n$ il peut se présenter trois types de situations :

1. On a effectivement observé A et B, et donc on peut en déduire une véritable observation Y_i de Y.

2. On a observé A, mais au moment où l'on considère l'échantillon, B n'a pas encore été observé pour certains éléments du n-échantillon. Tout ce que l'on sait est que $Y_i \geq C_i$, où C_i représente le moment de l'observation compté à partir de A. On est en présence d'une « *censure droite* ».

3. L'événement A a aussi été précédement observé, mais au moment où l'on considère l'échantillon, B a eu lieu sans qu'il soit lui-même observé. On est cette fois en présence d'une « *censure gauche* ». Tout ce que l'on sait est que $Y_i \leq C_i$.

Il existe d'autres types de censures, dues, par exemple, à la non-observation de A, ou encore des censures droites et gauches simultanées.

Nous allons maintenant préciser ce que l'on entend par censure, ne serait-ce que pour simuler le phénomène à l'aide d'un programme numérique. Nous n'envisagerons ici que la censure aléatoire à droite : c'est dire que nous nous plaçons dans le cas de « *censure droite* » précédent, où les C_i sont des variables aléatoires indépendantes des Y_j.

A partir de certaines observations, on a construit un n-échantillon ($T_1, \ldots,$-T_n), composé de données censurées et non-censurées. S'il n'y avait pas de censure, on observerait un autre n-échantillon $(Y_1, \ldots, Y_n)$ issu de la population suivant la loi F que l'on cherche à estimer. L'individu i a été observé au temps C_i, qui est aussi une variable aléatoire, mais qui suit la loi G. En effet il n'y a aucune raison pour que la loi suivie par le phénomène et celle qui dicte nos moments d'observations soient identiques. En cas de censure aléatoire à droite, seuls deux cas sont possibles : ou bien on a observé avant l'apparition de B, auquel cas l'observation est censurée à droite, ou bien on a effectivement observé B. On a alors :

$$T_i = \min(Y_i, C_i)\,. \tag{16.12}$$

Il est pratique d'introduire l'indicatrice D_i qui dit si l'observation est censurée ou non :

$$D_i = \mathbf{1}_{Y_i \leq C_i}\,. \tag{16.13}$$

16.2.2 L'estimateur de Kaplan-Meier.

Nous allons employer une terminologie classique issue de l'étude des durées de vie. Suivant cette étude, on cherche à estimer la fonction de survie $S(t) = 1 - F(t) \equiv \Pr\{T > t\}$. L'événement A est l'entrée d'un individu dans un flux

d'événements possédant la propriété étudiée, l'événement B est la sortie de l'individu de ce flux. Le temps que l'individu passe dans le flux est sa « *durée de vie* ». A chaque instant t, on a affaire à deux sortes d'individus :

1. Les individus qui sortent au temps t : ils sont en nombre $M(t)$. Ce nombre n'est différent de zéro qu'aux instants de sorties observées.

2. Les individus « restants » encore appelés individus « à risque », sont ceux qui au temps t ne sont ni sortis ni censurés avant le temps t : ils sont en nombre $R(t)$. Ce sont les sortants potentiels, les autres sont soit sortis, soit perdus de vue.

En utilisant la variable aléatoire D_i définie plus haut et qui vaut donc 0 si l'observation est censurée et 1 si elle ne l'est pas, on a :

$$M(t) = \sum_i D_i \mathbf{1}_{T_i=t} \,, \tag{16.14}$$

$$R(t) = \sum_i \mathbf{1}_{T_i \geq t} \,. \tag{16.15}$$

Avec cette notation, l'estimateur de Kaplan-Meier pour $S(t)$ est donné par la formule (Kaplan, Meier, (1958) [39]) :

$$\hat{S}_{\mathrm{KM}}(t) = \prod_{T_i \leq t} \left(1 - \frac{M(T_i)}{R(T_i)}\right) \,. \tag{16.16}$$

Donnons quelques propriétés de cet estimateur. Pour une discussion plus complète se reporter à Droesbeke *et al.*, (1989) [19].

- Sous des hypothèses assez générales, $\hat{S}_{\mathrm{KM}}$ converge presque-sûrement uniformément vers S. C'est l'équivalent du théorème de Glivenko-Cantelli en présence de censure.

- La fonction $\hat{S}_{\mathrm{KM}}$ ne présente de points de discontinuité qu'aux instants de sorties observées. Entre ces instants, cette fonction reste constante.

- La discontinuité tient compte des individus censurés, mais au point de discontinuité suivant où ils sont considérés comme étant sortis, ils ne font plus partie du contingent à risque. La fonction $\hat{S}_{\mathrm{KM}}$ est alors multipliée par un facteur qui est égal à la proportion des individus à risque qui ne sont pas sortis pendant l'intervalle précédent (1−la proportion de ceux qui sortent maintenant).

Pour des applications de l'estimateur de Kaplan-Meier au domaine de l'astronomie voir, par exemple, Feigelson et Nelson (1985) [20] ou Schmitt (1985) [65].

16.3 Densité de probabilité empirique.

Si l'on calcule la dérivée (au sens des distributions) de F_n, on obtient une somme de fonctions de Dirac :

$$\frac{d}{dx} F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \delta(x - X_i) \,. \tag{16.17}$$

Cette densité de probabilité peut être considérée comme un estimateur de la densité $f = dF/dx$, quand elle existe. Cette estimation contient le principe sous-jacent menant aux méthodes « bootstrap ». Cependant la fonction ainsi calculée ne possède généralement pas les propriétés connues de la densité f. Il est possible, par exemple, que l'on sache que f est continue et il serait alors souhaitable de l'estimer par une fonction elle aussi continue, ce qui n'est pas le cas de l'estimateur (16.17) ci-dessus.

16.3.1 Estimateurs subordonnés à un noyau.

Pour résoudre ce problème, on a pensé (Rosenblatt (1956) [63], Parzen (1962) [55]) à convoluer les fonctions δ par un noyau $K(x)$ afin d'obtenir les estimateurs :

$$\widehat{f}_{K,n}(x) = \frac{1}{nh(n)} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h(n)}\right) . \tag{16.18}$$

Le paramètre h contrôle le degré de « lissage » appliqué aux fonctions δ : on l'appelle souvent la « *fenêtre de lissage* ». On a démontré que si le noyau $K(x)$ et le paramètre h possèdent les propriétés suivantes :

$$\lim_{|x|\to\infty} |x| K(x) = 0, \quad \int_{-\infty}^{\infty} K(x)dx = 1, \tag{16.19}$$

$$\lim_{n\to\infty} h(n) = 0, \quad \lim_{n\to\infty} \frac{1}{nh(n)} = 0, \tag{16.20}$$

alors l'estimateur $\widehat{f}_{K,n}$ est convergent et asymptotiquement non-biaisé. On impose en général la condition $0 < K(x) < \infty$, mais elle n'est pas nécessaire. On peut être guidé dans le choix de K par des considérations sur le comportement à l'infini des « ailes » de la densité f. Un noyau classique est :

$$K(x) = \frac{1}{\sqrt{2\pi}} \exp -\frac{x^2}{2}, \tag{16.21}$$

qui, pour $h(n) = \sigma/\sqrt{n}$, conduit à l'estimateur :

$$\widehat{f} = \frac{1}{\sqrt{2\pi n}\sigma} \sum_{i=1}^{n} \exp -\frac{(x - X_i)^2}{2\sigma^2/n} . \tag{16.22}$$

Pour une application de ces estimateurs dans le domaine de l'astronomie voir de Jager *et al.* (1986) [34].

16.4 Caractéristiques numériques de la loi empirique.

La loi empirique admettant F_n pour fonction de répartition possède des caractéristiques numériques que l'on calcule comme espérance mathématique d'une fonction de la variable aléatoire X. Ainsi, pour les moments, on a :

$$\mathrm{E}\left\{X^k\right\} = \int x^k \, dF_n . \tag{16.23}$$

Il est facile de montrer que les moments $\mathrm{E}\{X^k\}$ sont égaux aux moments empiriques M'_k. On a en effet :

$$\mathrm{E}\left\{X^k\right\} = \frac{1}{n}\sum_{i=1}^{n} X_i = M'_k\,. \tag{16.24}$$

Les caractéristiques numériques ainsi obtenues peuvent servir d'estimateurs des caractéristiques numériques de la loi F.

16.5 Histogrammes.

On a encore l'habitude de regouper les variables aléatoires $X_{(i)}$, dans des cellules de même largeur $h = \Delta x$ et d'en compter le nombre de façon à obtenir une courbe discontinue en «escalier» appelée «*histogramme*». Plus précisément, on choisit un intervalle entre deux valeurs extrêmes $x_{\min}$ et $x_{\max}$, que l'on découpe en k cellules de largeur $\Delta x = (x_{\max} - x_{\min})/k$. A partir du n-échantillon $(X_1, \ldots, X_n)$, on fabrique les $k+2$ variables aléatoires P_i telles que :

$$x_i = x_{\min} + (i-1)\Delta x, \tag{16.25}$$

$$P_i = \int_{x_i}^{x_i+\Delta x} \frac{1}{n}\sum_{j=1}^{n} \delta(x - X_{(j)})\,dx, \quad i = 1, \ldots, k, \tag{16.26}$$

$$P_0 = \int_{-\infty}^{x_{\min}} \frac{1}{n}\sum_{j=1}^{n} \delta(x - X_{(j)})\,dx, \tag{16.27}$$

$$P_{k+1} = \int_{x_{\max}}^{\infty} \frac{1}{n}\sum_{j=1}^{n} \delta(x - X_{(j)})\,dx\,. \tag{16.28}$$

Ces variables aléatoires P_i représentent le nombre de X_i qui se trouvent dans la i-ème cellule de l'histogramme, divisé par la taille n de l'échantillon. L'histogramme de $(X_1, \ldots, X_n)$ contient moins d'information que la fonction de répartition empirique F_n, car on perd l'ordre des $X_{(i)}$ dans une cellule.

16.5.1 Loi suivie par le nombre de points dans une cellule.

Le nombre N_i de variables aléatoires telles que X_i ou, plus rapidement, le nombre de points dans la cellule i, est par définition égal à nP_i. Ce nombre est également une variable aléatoire dont la loi dépend du fait que le nombre total de points n est ou non une variable aléatoire.

Cas où n n'est pas une variable aléatoire.

Le nombre de points à répartir dans les $k+2$ cellules est connu à l'avance. La probabilité pour qu'un point tombe dans la cellule numéro i est donnée par :

$$p_i = \Pr\{x_i < X_i \leq x_i + \Delta x\} = \int_{x_i}^{x_i+\Delta x} f(x)\,dx, \tag{16.29}$$

pour $i = 1, \ldots, k$ et, pour $i = 0, k+1$, par :

$$p_0 = \int_{-\infty}^{x_{\min}} f(x)\,dx, \quad p_{k+1} = \int_{x_{\max}}^{\infty} f(x)\,dx\,. \tag{16.30}$$

La répartion des points sur l'axe des x s'effectuant de façon indépendante, on obtient finalement une loi binomiale :

$$\Pr\{N_i = n_i\} = C_n^{n_i} p_i^{n_i} (1-p_i)^{n-n_i} \ . \tag{16.31}$$

La moyenne et la variance du nombre de point N_i dans la cellule i d'un histogramme valent donc :

$$\mathrm{E}\{N_i\} = np_i, \quad \mathrm{Var}(N_i) = np_i(1-p_i) \ . \tag{16.32}$$

L'ensemble des points de l'histogramme suit alors une loi multinomiale d'expression :

$$\Pr\{N_0 = n_0, N_1 = n_1, \ldots, N_{k+1} = n_{k+1}\} = \frac{n!}{n_0! n_1! \ldots n_{k+1}!} p_0^{n_0} p_1^{n_1} \ldots p_{k+1}^{n_{k+1}} \tag{16.33}$$

En particulier, les variables aléatoires N_i et $N_j, i \neq j$ sont corrélées, avec pour coefficient de corrélation :

$$\rho_{ij} = -\sqrt{\frac{p_i p_j}{(1-p_i)(1-p_j)}} \ . \tag{16.34}$$

Le coefficient de corrélation ρ_{ij} étant négatif, ces variables aléatoires sont en fait anti-corrélées, ce qui indique que, par exemple, N_i tend à diminuer lorsque N_j augmente. Ce comportement est naturel puisque le nombre de points à répartir est fixé.

Cas où n est une variable aléatoire.

Si le nombre de points observés n n'est pas déterminé à l'avance, mais est une variable aléatoire N, on trouvera la loi suivie par la variable aléatoire N_i comme somme des probabilités conditionnelles suivantes :

$$\Pr\{N_1 = n_1\} = \sum_{k=n_1}^{\infty} \Pr\{N_1 = n_1 | N = k\} \Pr\{N = k\} \ . \tag{16.35}$$

En particulier si le nombre de points à distribuer dans les cellules de l'histogramme suit une loi de Poisson de paramètre μ tel que $\mathrm{E}\{N\} = \mu$, on aura :

$$\Pr\{N_i = n_i\} = \sum_{k=n_i}^{\infty} C_k^{n_i} p_i^{n_i} (1-p_i)^{k-n_i} \frac{\mu^k}{k!} e^{-\mu} \ . \tag{16.36}$$

Un calcul simple montre que finalement :

$$\Pr\{N_i = n_i\} = \frac{(\mu p_i)^{n_i}}{n_i!} e^{-\mu p_i}, \tag{16.37}$$

où μ est le nombre moyen de points dans l'histogramme et p_i la probabilité pour qu'un point « tombe » dans la i-ème cellule. Le nombre de points N_i dans la cellule numéro i est donc, dans ce cas, une variable aléatoire de Poisson ayant pour moyenne et pour variance :

$$\mathrm{E}\{N_i\} = \mathrm{E}\{N\} p_i, \quad \mathrm{Var}(N_i) = \mathrm{E}\{N\} p_i \ . \tag{16.38}$$

Les variables aléatoires N_i et $N_j, i \neq j$ sont indépendantes.

16.5.2 Le χ^2 de Pearson.

Pearson a montré que la statistique X^2 définie par l'expression ci-dessous :

$$X^2 = \sum_{i=0}^{k+1} \frac{(N_i - Np_i)^2}{Np_i} \tag{16.39}$$

est une variable aléatoire tendant vers une loi du χ^2 à $k+1$ degrés de liberté quand $N = n$ est connu à l'avance et tend vers l'infini, et à $k+2$ degrés de liberté quand N est une variable aléatoire dont la moyenne tend vers l'infini. Rappelons que $k+2$ est le nombre de cellules de l'histogramme. La statistique X^2 est utilisée dans le test du χ^2 qui vise à décider de la conformité de l'échantillon $(X_1, \ldots, X_n)$ qui a servi à construire l'histogramme vis-à-vis de la loi F qui a permis de calculer les p_i.

16.5.3 Taille des cellules.

La raison qui pousse à construire des histogrammes correspond à un souci de réduction des données. La taille Δx ou le nombre de cellules $k+2$ est fixé à partir d'un compromis entre les deux objectifs contradictoires suivants : cette taille ne doit pas être trop grande afin de ne pas trop perdre l'information sur l'ordre des points, elle ne doit pas être trop petite non plus, auquel cas les cellules sont presque toutes vides. Concrètement, le meilleur Δx sera celui qui minimisera une erreur commise lorsque l'on approxime la densité f par l'histogramme $\widehat{f}_{\Delta x}$. La taille des cellules peut donc être implicitement définie par :

$$\min_{\Delta x} \int_{x_{\min}}^{x_{\max}} |\widehat{f}_{\Delta x}(x) - f(x)|^2 \, dx \ . \tag{16.40}$$

Un bon choix de Δx est alors :

$$\Delta x^* = \left[6 / \left((n - n_{\text{out}}) \int_{x_{\min}}^{x_{\max}} |f'(x)|^2 \, dx \right) \right]^{\frac{1}{3}} \tag{16.41}$$

où n_{out} est le nombre de points en dehors de l'intervalle $[x_{\min}, x_{\max}]$.

Chapitre 17

Etude de la dépendance.

On présente souvent les résultats d'une expérience comportant plusieurs observations par un ensemble de points dispersés sur un plan rapporté à un repère orthonormé xOy. Chaque observation i fournit deux nombres X_i et Y_i, que l'on considère alors comme les coordonnées d'un point P_i dans ce repère xOy. En général, les points se répartissent dans ce plan sous la forme d'un « nuage » et ne se regroupent pas sur une courbe telle que $\varphi(x, y) = 0$. On interprète ce phénomène comme dû au fait qu'une au moins des coordonnées X_i et/ou Y_i, est une variable aléatoire. Chaque mesure, en outre, est susceptible d'être entachée d'erreur. On envisage alors, selon qu'une ou les deux coordonnées sont aléatoires, ou selon qu'il y a présence de bruit ou non, diverses méthodes d'analyse des données. On distingue couramment l'étude de la corrélation, de la régression, et la recherche de dépendance fonctionnelle.

17.1 Etude de la corrélation.

Suivant ce modèle, on considère les n observations comme le résultat d'un échantillonnage dans une population décrite par une loi F à deux dimensions qui est indépendante de l'indice i du « tirage ». Les variables aléatoires X_i et Y_i suivent alors quel que soit i les mêmes lois, qui sont les lois marginales de F. Ces variables aléatoires peuvent être indépendantes ou dépendantes. Une mesure de la dépendance affine est le coefficient de corrélation ρ, défini par :

$$\rho(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}} \cdot \tag{17.1}$$

Lorsque $\rho \to \pm 1$, les points (X_i, Y_i) tendent à se regrouper de façon à satisfaire la relation affine :

$$\frac{Y - \mathrm{E}\{Y\}}{\sqrt{\mathrm{Var}(Y)}} = \rho \frac{X - \mathrm{E}\{X\}}{\sqrt{\mathrm{Var}(X)}} \cdot \tag{17.2}$$

L'étude de la corrélation vise à estimer le coefficient de corrélation ρ à partir d'un échantillon de taille n, issu de la population 2D, puis à conclure à une plus ou moins grande dépendance affine entre les variables aléatoires X et Y, suivant

que $|\rho|$ est proche de 1 ou de 0. Avant de poursuivre, il faut bien prendre garde aux points suivants :

- Si l'on conclut que $\rho = 0$, cela signifie que les X_i et Y_i sont non-corrélés, mais cela ne signifie surtout pas que ces variables aléatoires sont indépendantes. En revanche, des variables aléatoires normales non-corrélées sont indépendantes.

- S'il ressort de l'analyse de la corrélation que $|\rho|$ est proche de 1, cela indique une forte tendance affine entre X_i et Y_i. Par exemple si ρ est proche de 1, X_i et Y_i auront tendance à augmenter ou diminuer en même temps, et en sens contraire si ρ est proche de -1. Mais cela n'implique pas nécessairement qu'il existe une relation de cause à effet entre les deux variables.

17.1.1 Coefficient de corrélation en présence d'erreurs de mesure.

Les variables aléatoires X et Y peuvent être sujettes à des erreurs de mesures, et ces erreurs ont pour effet de modifier le coefficient de corrélation ρ du couple (X, Y). Si, par exemple, viennent s'ajouter à X et à Y des erreurs U et V de moyenne nulle, de variances σ_U^2, σ_V^2 non corrélées entre elles et non corrélées avec les X, Y, alors le coefficient de corrélation des mesures ρ^* est le coefficient de corrélation du couple $(X + U, Y + V)$ qui est donné par l'expression :

$$\rho^* = \rho \left[\left(1 + \frac{\sigma_U^2}{\sigma_X^2} \right) \left(1 + \frac{\sigma_V^2}{\sigma_Y^2} \right) \right]^{-\frac{1}{2}} . \tag{17.3}$$

L'effet des erreurs de mesures non-corrélées est de diminuer la valeur du coefficient de corrélation.

17.1.2 L'estimateur « naturel » de ρ.

En remplaçant les moments de la population par les moments de l'échantillon dans (17.1), on obtient un estimateur R de ρ dit « estimateur naturel » :

$$R = \frac{\frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2 \frac{1}{n} \sum_{i=1}^{n} (Y_i - \bar{Y})^2}} . \tag{17.4}$$

Comme ρ, R est compris entre -1 et 1. Nous appellerons R, le « *coefficient de corrélation empirique* », de l'échantillon. C'est une variable aléatoire que nous allons maintenant étudier.

17.1.3 Le cas normal.

Si la loi parente 2D est normale, sa densité est alors donnée par l'expression :

$$f(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y(1-\rho^2)^{\frac{1}{2}}} \exp\left\{-\frac{1}{2(1-\rho^2)}\right.$$

$$\left.\left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right]\right\}, \quad (17.5)$$

où μ_X, μ_Y sont les moyennes de X et Y, σ_X, σ_Y leurs écart types et ρ leur coefficient de corrélation.

La densité de probabilité d'un n-échantillon $(X_1, \ldots, X_n)$ indépendant (i.i.d) est donnée par le produit des densités simples :

$$f_n(x_1, y_1, \ldots, x_n, y_n) = \prod_{i=1}^{n} f(x_i, y_i)\,. \quad (17.6)$$

La fonction de répartition de la loi suivie par R est trouvée en intégrant dans l'espace de définition du n-échantillon sur tout le domaine où R est inférieur à un certain seuil r :

$$F_R(r|\rho) = \int \cdots \int_{R\leq r} f_n(x_1, y_1, \ldots, x_n, y_n)\, dx_1 dy_1 \cdots dx_n dy_n\,. \quad (17.7)$$

La densité de probabilité f_R de R est, par définition, la dérivée de F_R. On a :

$$F_R(r|\rho) = \int_{-1}^{r} f_R(u|\rho)\, du, \qquad f_R(r|\rho) = \frac{d}{dr} F_R(r)\,. \quad (17.8)$$

Fisher (1915) [21] a donné l'expression de cette densité :

$$f_R(r|\rho) = \frac{(1-\rho^2)^{\frac{n-1}{2}}}{\pi(n-3)!}(1-r^2)^{\frac{n-4}{2}} \frac{d^{n-2}}{d(r\rho)^{n-2}} \frac{\arccos(-r\rho)}{\sqrt{1-r^2\rho^2}}\,. \quad (17.9)$$

Une forme numériquement plus maniable est due à Hotelling (1953) [33] :

$$f_R(r|\rho) = \frac{n-2}{\sqrt{2}(n-1)B(\frac{1}{2}, n-\frac{1}{2})}(1-\rho^2)^{\frac{1}{2}(n-1)}(1-r^2)^{\frac{1}{2}(n-4)}(1-r\rho)^{\frac{3}{2}-n}$$
$$\times F(\tfrac{1}{2}, \tfrac{1}{2}, n-\tfrac{1}{2}, \tfrac{1}{2}(1+r\rho))\,. \quad (17.10)$$

Dans cette dernière expression, B est la fonction eulérienne de première espèce, et F est la fonction hypergéométrique.

La loi suivie par R possède une moyenne et une variance données par :

$$\mathrm{E}\{R|\rho\} = \rho\left[1 - \frac{(1-\rho^2)^2}{2n} + O(n^{-2})\right], \quad (17.11)$$

$$\mathrm{Var}(R|\rho) = \frac{(1-\rho)^2}{n-1}\left(1 + \frac{11\rho^2}{2n}\right) + O(n^{-3})\,. \quad (17.12)$$

Les coefficients d'asymétrie et d'aplatissement sont donnés par :

$$\gamma_1 = -\frac{6\rho}{\sqrt{n}} + o(n^{\frac{1}{2}}), \quad \gamma_2 = \frac{6(12\rho^2 - 1)}{n} + o(n^{-1}) \ . \tag{17.13}$$

L'expression (17.11) montre que R est un estimateur légèrement biaisé de ρ ; Olkin et Pratt (1958) [52] ont trouvé l'estimateur, par ailleurs unique, R_0 non-biaisé de ρ :

$$R_0 = F(\tfrac{1}{2}, \tfrac{1}{2}, \tfrac{1}{2}(n-2), (1-R^2))R \ . \tag{17.14}$$

De façon pratique, le biais en $1/n$ est retiré lorsque l'on utilise la formule approchée :

$$R_0 \approx \left(1 + \frac{1-R^2}{2(n-4)}\right) R \ \cdot \tag{17.15}$$

L'expression (17.12) montre que R est un estimateur convergent de ρ, et par conséquent R_0 est également convergent. Dans le cas particulier où $\rho = 0$, la densité de probabilité de R devient :

$$\boxed{f_R(r|\rho = 0) = \frac{1}{B(\frac{1}{2}, \frac{1}{2}(n-2))}(1-r^2)^{\frac{1}{2}(n-4)} \ .} \tag{17.16}$$

Dans ce cas on a :

$$\mathrm{E}\{R|\rho = 0\} = 0, \quad \mathrm{Var}(R|\rho = 0) = \frac{1}{n-1}, \tag{17.17}$$

et l'estimateur R est alors non-biaisé. Comme le montre la figure 17.1, les densités de probabilité f_R tendent très lentement vers la loi normale. Notons que pour un échantillon de taille 4 issu d'une population où $\rho = 0$, R est distribué uniformément entre -1 et 1.

17.1.4 Estimation d'intervalle.

Les expressions de f_R permettent de calculer des abaques donnant l'estimation d'intervalle suivant la méthode exposée au chapitre « Estimation d'intervalle ». Ces abaques servent à trouver un intervalle $]\rho_{\min}, \rho_{\max}]$ qui, pour un coefficient de confiance γ donné, satisfait l'équation :

$$\Pr\{\rho_{\min}(R) < \rho \leq \rho_{\max}(R)\} = \gamma \ . \tag{17.18}$$

Les valeurs $\rho_{\min}$ et $\rho_{\max}$, pour $\gamma = 1 - 2\alpha$ donné, sont solutions de :

$$1 - \alpha = F_R(r|\rho_{\min}), \quad \alpha = F_R(r|\rho_{\max}). \tag{17.19}$$

Dans ce cas, il s'agit l'intervalle bilatéral symétrique.

17.2 La régression.

L'échantillon sur lequel va porter l'étude de la régression est toujours un échantillon indépendant et identiquement distribué (i.i.d), issu d'une loi à deux

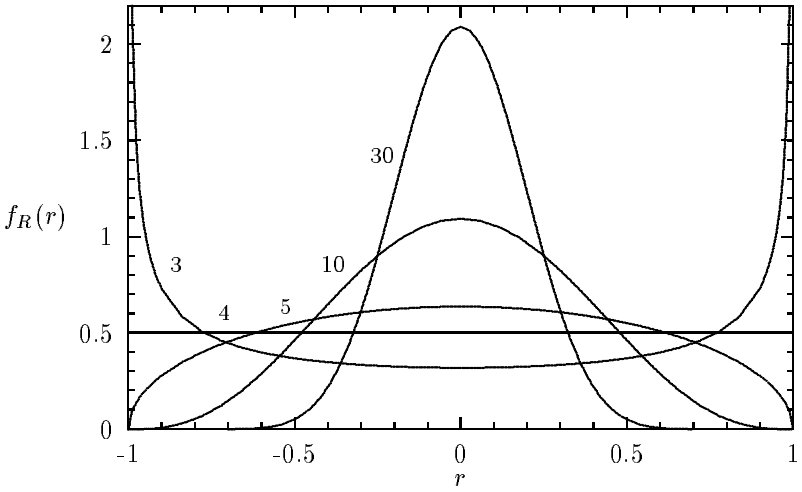


FIG. 17.1: *Exemples de densités de lois suivies par le coefficient de corrélation empirique R quand la population est non-corrélée ($\rho = 0$) et pour des échantillons de tailles 3,4,5,10 et 30. Il faut remarquer que lorsque la taille de l'échantillon est égale à 4, on trouve R avec une densité uniforme entre -1 et 1.*

dimensions. On s'intéresse maintenant aux lois conditionnelles et plus spécialement aux lois suivies par une variable, sachant que l'autre est connue. Ces lois possèdent, le plus souvent, des moyennes conditionnelles η ainsi définies :

$$\eta_{Y|X} = \mathrm{E}\left\{Y|X=x\right\}, \quad \eta_{X|Y} = \mathrm{E}\left\{X|Y=y\right\} \ . \tag{17.20}$$

La moyenne des Y, sachant que $X = x$, $\eta_{Y|X}$, est en général une fonction de x que l'on appelle « *courbe de régression de Y par rapport à X* », et de façon similaire, la moyenne des X sachant que $Y = y$ est appelée « *courbe de régression de X par rapport à Y* ».

17.2.1 La régression linéaire.

Certaines lois possèdent des courbes de régression à dépendance affine que, par abus de langage, on appelle « linéaire » . Pour ces lois, les paramètres de la régression α_1, α_2 et β_1, β_2 sont par définition :

$$\mathrm{E}\left\{Y|X=x\right\} = \alpha_2 + \beta_2 x, \quad \mathrm{E}\left\{X|Y=y\right\} = \alpha_1 + \beta_1 y \ . \tag{17.21}$$

C'est, par exemple, le cas de la loi normale 2D de paramètres $\mu_1, \mu_2, \sigma_1, \sigma_2$ et ρ. Cherchons maintenant les relations pouvant exister entre les coefficients α, β de la régression et les caractéristiques numériques de la loi parente, quand elles existent. Montrons tout d'abord que les droites de régression se coupent au point de coordonnées (μ_1, μ_2), correspondant à la moyenne de la loi. En effet, si l'on pose $x = \mu_1$ dans (17.21), on aura $\mathrm{E}\{Y|X=\mu_1\} = \alpha_2 + \beta_2\mu_1$, mais $\mathrm{E}\{Y|X=\mu_1\} = \mu_2$. En posant de même $y = \mu_2$ pour la droite de régression suivant X on trouverait finalement :

$$\mu_2 = \alpha_2 + \beta_2\mu_1, \quad \mu_1 = \alpha_1 + \beta_1\mu_2 \ . \tag{17.22}$$

De façon imagée, cela exprime que les droites de régression se coupent au centre de gravité de la loi parente. On aurait trouvé le même résultat en raisonnant de la façon suivante : si x est une variable aléatoire suivant la loi marginale en x du couple (X, Y), on a :

$$\mu_2 = \mathrm{E}\{Y\} = \mathrm{E}\{\mathrm{E}\{Y|x = X\}\} = \mathrm{E}\{\alpha_2 + \beta_2 X\} = \alpha_2 + \beta_2\mu_1,$$

et un résultat similaire pour l'autre droite de régression. En suivant un raisonnement analogue, on démontre alors à l'aide de la covariance μ_{11} :

$$\begin{aligned}\mu_{11} &= \mathrm{E}\{(X-\mu_1)(Y-\mu_2)\} = \mathrm{E}\{\mathrm{E}\{(X-\mu_1)(Y-\mu_2)|x=X\}\}\ , \\ &= \mathrm{E}\{(X-\mu_1)\,\mathrm{E}\{(Y-\mu_2)|x=X\}\}\ .\end{aligned}$$

Evaluons l'espérance la plus interne en utilisant (17.21) et (17.22). Il vient :

$$\begin{aligned}\mathrm{E}\{(Y-\mu_2)|x=X\} &= \mathrm{E}\{Y|x=X\} - \mu_2\,, \\ &= \alpha_2 + \beta_2 X - \mu_2\,, \\ &= \alpha_2 + \beta_2 X - \alpha_2 - \beta_2\mu_1\,, \\ &= \beta_2(X-\mu_1)\,,\end{aligned}$$

soit finalement :

$$\mu_{11} = \mathrm{E}\left\{\beta_2(X-\mu_1)^2\right\} = \beta_2\sigma_1^2\,. \tag{17.23}$$

De façon analogue on montrerait que :

$$\mu_{11} = \mathrm{E}\left\{\beta_1(Y-\mu_2)^2\right\} = \beta_1\sigma_2^2, \tag{17.24}$$

d'où $\mu_{11} = \beta_1\sigma_2^2 = \beta_2\sigma_1^2$. En introduisant le coefficient de corrélation $\rho = \mu_{11}/(\sigma_1\sigma_2)$, il vient :

$$\rho^2 = \beta_1\beta_2\ . \tag{17.25}$$

Cette dernière propriété, en conjonction avec (17.23) et (17.24), montre que les droites de régression sont confondues si et seulement si $|\rho| = 1$ et qu'elles sont orthogonales et parallèles aux axes si et seulement si les variables aléatoires X et Y sont non-corrélées, cela dans le cadre des lois non-dégénérées où $\sigma_1, \sigma_2 \neq 0$.

Finalement on tire de (17.22), (17.23) et (17.24) l'expression des coefficients de la régression :

$$\alpha_2 = \mu_2 - \rho\frac{\sigma_2}{\sigma_1}\mu_1, \quad \beta_2 = \rho\frac{\sigma_2}{\sigma_1}\,, \tag{17.26}$$

$$\alpha_1 = \mu_1 - \rho\frac{\sigma_1}{\sigma_2}\mu_2, \quad \beta_1 = \rho\frac{\sigma_1}{\sigma_2}\,, \tag{17.27}$$

soit, en remplaçant dans la définition (17.21) des droites de régression :

$$\mathrm{E}\{Y|X=x\} = \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x-\mu_1)\,, \tag{17.28}$$

$$\mathrm{E}\{X|Y=y\} = \mu_1 + \rho\frac{\sigma_1}{\sigma_2}(y-\mu_2)\,. \tag{17.29}$$

17.2.2 Droites de régression empiriques.

A partir d'un n-échantillon i.i.d, on peut calculer les droites de régression empiriques en remplaçant les moments de la loi dans les expressions (17.26) et (17.27), par les moments empiriques, ce qui conduit aux estimateurs :

$$\widehat{\beta}_2 = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}, \quad \bar{Y} = \widehat{\alpha}_2 + \widehat{\beta}_2 \bar{X}, \tag{17.30a}$$

$$\widehat{\beta}_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2}, \quad \bar{X} = \widehat{\alpha}_1 + \widehat{\beta}_1 \bar{Y}. \tag{17.30b}$$

Ces expressions sont identiques à celles que l'on aurait trouvées si l'on avait calculé les coefficients des droites de moindres carrés, respectivement suivant le modèle $y = \alpha_2 + \beta_2 x$ et suivant le modèle $x = \alpha_1 + \beta_1 y$, (voir les formules (14.97)).

17.3 Recherche de dépendances fonctionnelles.

On suppose maintenant que le nuage de points (X_i, Y_i) est dû au déplacement aléatoire $(\Delta X_i, \Delta Y_i)$ de ces points à partir du point $(\dot{x}_i, \dot{y}_i)$, voir figure 17.2. La

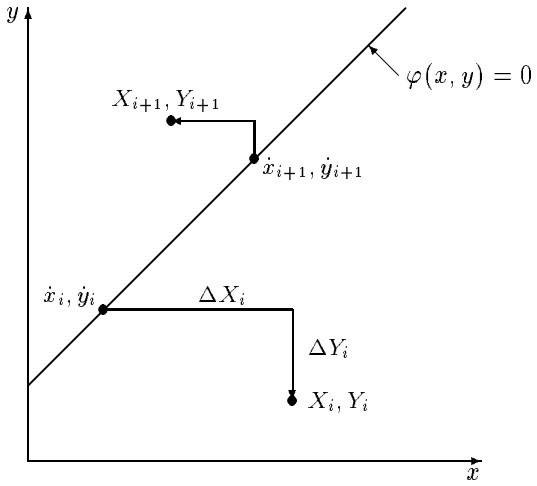


FIG. 17.2: *Schéma de principe de la recherche d'une dépendance fonctionnelle.*

cause de ce déplacement aléatoire est la présence de bruit dans les mesures visant à déterminer $(\dot{x}_i, \dot{y}_i)$. Supposons en outre que l'on ait de bonnes raisons de croire que les points $(\dot{x}_i, \dot{y}_i)$ obéissent à la relation fonctionnelle $\varphi(x, y) = 0$, où la fonction φ est connue à un certain nombre de paramètres θ_k près. Une autre façon de présenter les choses est de dire que si les erreurs de mesures tendaient vers 0, alors les points expérimentaux se regrouperaient le long de la courbe $\varphi(x, y) = 0$.

Si la loi qui préside à l'apparition des erreurs de mesures possède une moyenne, il est naturel d'identifier le point $(\dot{x}_i, \dot{y}_i)$ à cette moyenne. Le problème qui se

pose alors est de trouver des estimateurs $\widehat{\theta}_k$ des θ_k, à partir d'un échantillon de taille n $(X_i, Y_i), i = 1, \ldots, n$. Le logarithme de la fonction de vraisemblance de l'échantillon est donnée par l'expression :

$$L(\theta_1, \ldots, \theta_k) = \ln f_{\text{ech}}(x_1, y_1, \ldots, x_n, y_n | \theta_1, \ldots, \theta_k), \tag{17.31}$$

où f_{ech} est la densité de la loi suivie par l'échantillon. Le calcul est simplifié si les erreurs de mesures sont indépendantes entre elles. On a alors :

$$L = \sum_{i=1}^{n} \ln f_i(x_i, y_i | \theta_1, \ldots, \theta_k), \tag{17.32}$$

où f_i est la densité de la loi suivie par le point (X_i, Y_i). Calculons cette expression dans le cas où les erreurs de mesures ΔX_i et ΔY_i suivent une loi normale 2D de moyenne $(\dot{x}_i, \dot{y}_i)$, de variance $(\sigma_{xi}, \sigma_{yi})$ et de coefficient de corrélation $\rho = 0$. Il vient, à une constante additive près :

$$L = -\frac{1}{2} \sum_{i=1}^{n} \frac{(x_i - \dot{x}_i)^2}{\sigma_{xi}^2} + \frac{(y_i - \dot{y}_i)^2}{\sigma_{yi}^2} \,. \tag{17.33}$$

Dans cette expression les couples (x_i, y_i) représentent une réalisation des (X_i, Y_i). Les valeurs $(\dot{x}_i, \dot{y}_i)$ que l'on cherche sont soumises à la contrainte $\varphi(\dot{x}_i, \dot{y}_i | \theta_1, \ldots, \theta_k) = 0$. Le principe du maximum de vraisemblance nous prescrit de choisir, en tant qu'estimation des $(\dot{x}_i, \dot{y}_i)$, les valeurs $(\widehat{x}_i, \widehat{y}_i)$ qui rendent l'expression (17.33) maximum. Finalement le problème à résoudre est de trouver les $(\widehat{x}_i, \widehat{y}_i), i = 1, \ldots, n$ et les $\widehat{\theta}_j, j = 1, \ldots, k$, tels que :

$$L(\widehat{x}_1, \widehat{y}_1, \ldots, \widehat{x}_n, \widehat{y}_n) = \max_{\tilde{x}_i, \tilde{y}_i} -\frac{1}{2} \sum_{i=1}^{n} \frac{(x_i - \tilde{x}_i)^2}{\sigma_{xi}^2} + \frac{(y_i - \tilde{y}_i)^2}{\sigma_{yi}^2}, \tag{17.34}$$

sujets aux n contraintes :

$$\varphi_i(\widehat{x}_i, \widehat{y}_i | \widehat{\theta}_1, \ldots, \widehat{\theta}_k) = 0, \quad i = 1, \ldots, n \,. \tag{17.35}$$

La méthode classique pour résoudre ce type de problème est la méthode des multiplicateurs de Lagrange. Nous allons illustrer son fonctionnement dans le cas où φ est affine (voir également York (1966) [72]) :

$$\varphi(x, y) = y - ax - b = 0 \,. \tag{17.36}$$

$$\varphi(\widehat{x}_i, \widehat{y}_i) = \widehat{y}_i - \widehat{a}\widehat{x}_i - \widehat{b} = 0 \quad i = 1, \ldots, n \,. \tag{17.37}$$

Ce modèle correspond au cas où l'on veut ajuster une droite dans un nuage de points, en tenant compte de l'existence d'erreurs suivant l'axe des x et suivant l'axe des y. Il vient :

$$\frac{\partial L}{\partial \widehat{x}_i} - \lambda_i \frac{\partial \varphi_i}{\partial \widehat{x}_i} = 0 \,, \qquad \frac{\partial L}{\partial \widehat{y}_i} - \lambda_i \frac{\partial \varphi_i}{\partial \widehat{y}_i} = 0 \,, \quad i = 1, \ldots, n \tag{17.38}$$

$$\frac{\partial L}{\partial \widehat{a}} - \sum_{i=1}^{n} \lambda_i \frac{\partial \varphi_i}{\partial \widehat{a}} = 0 \,, \qquad \frac{\partial L}{\partial \widehat{b}} - \sum_{i=1}^{n} \lambda_i \frac{\partial \varphi_i}{\partial \widehat{b}} = 0 \,. \tag{17.39}$$

Les λ_i sont les n multiplicateurs de Lagrange correspondant aux n contraintes. Ces équations expriment le fait que lorsque L est extremum, le gradient de L est une combinaison linéaire du gradient des φ_i. On a :

$$\frac{\partial L}{\partial \widehat{x}_i} = -\frac{x_i - \widehat{x}_i}{\sigma_{xi}^2}, \quad \frac{\partial L}{\partial \widehat{y}_i} = -\frac{y_i - \widehat{y}_i}{\sigma_{yi}^2}, \quad \frac{\partial L}{\partial \widehat{a}} = 0, \quad \frac{\partial L}{\partial \widehat{b}} = 0\,. \tag{17.40}$$

$$\frac{\partial \varphi_i}{\partial \widehat{x}_i} = -\widehat{a}, \quad \frac{\partial \varphi_i}{\partial \widehat{y}_i} = 1, \quad \frac{\partial \varphi_i}{\partial \widehat{a}} = -\widehat{x}_i, \quad \frac{\partial \varphi_i}{\partial \widehat{b}} = -1. \tag{17.41}$$

En remplaçant ces expressions dans (17.38), il vient :

$$-\frac{x_i - \widehat{x}_i}{\sigma_{xi}^2} - \lambda_i(-\widehat{a}) = 0\,, \tag{17.42}$$

$$-\frac{y_i - \widehat{y}_i}{\sigma_{yi}^2} - \lambda_i(+1) = 0\,, \tag{17.43}$$

$$\widehat{y}_i - \widehat{a}\widehat{x}_i - \widehat{b} = 0\,. \tag{17.44}$$

Supposons maintenant que les coefficients $\widehat{a}$ et $\widehat{b}$ soient connus. Le système à résoudre devient alors linéaire :

$$\begin{pmatrix} 0 & \sigma_{xi}^{-2} & \widehat{a} \\ \sigma_{yi}^{-2} & 0 & -1 \\ 1 & -\widehat{a} & 0 \end{pmatrix} \begin{pmatrix} \widehat{y}_i \\ \widehat{x}_i \\ \lambda_i \end{pmatrix} = \begin{pmatrix} \dfrac{x_i}{\sigma_{xi}^2} \\ \dfrac{y_i}{\sigma_{yi}^2} \\ \widehat{b} \end{pmatrix}, \tag{17.45}$$

et il a pour solution :

$$\widehat{y}_i = \frac{\widehat{a}^2\sigma_{xi}^2 y_i + \widehat{a}\sigma_{yi}^2 x_i + \widehat{b}\sigma_{yi}^2}{\sigma_{yi}^2 + \widehat{a}^2\sigma_{xi}^2}, \tag{17.46}$$

$$\widehat{x}_i = \frac{\widehat{a}\sigma_{xi}^2 y_i + \sigma_{yi}^2 x_i - \widehat{a}\widehat{b}\sigma_{yi}^2}{\sigma_{yi}^2 + \widehat{a}^2\sigma_{xi}^2}, \tag{17.47}$$

$$\lambda_i = \frac{y_i - \widehat{a}x_i - \widehat{b}}{\sigma_{yi}^2 + \widehat{a}^2\sigma_{xi}^2}. \tag{17.48}$$

Il est facile de montrer que $\widehat{x}_i$ et $\widehat{y}_i$ sont les coordonnées du point de contact avec la droite cherchée, d'une ellipse centrée en x_i, y_i et dont les axes sont dans le rapport σ_{xi}/σ_{yi}. La fonction de vraisemblance prend une forme plus simple si l'on introduit les poids w_i et les résidus z_i suivants :

$$w_i(a) = \frac{1}{\sigma_{yi}^2 + a^2\sigma_{xi}^2}, \quad z_i = y_i - (ax_i + b)\,. \tag{17.49}$$

On a alors :

$$L(x_1, y_1, \ldots, x_n, y_n | a, b) = \sum_{i=1}^{n} \frac{(y_i - ax_i - b)^2}{\sigma_{yi}^2 + a^2\sigma_{xi}^2} = \sum_{i=1}^{n} w_i z_i^2\,. \tag{17.50}$$

On a donc transformé un problème de minimisation avec contraintes, en un problème de minimisation pure. Il est possible d'aller plus loin en utilisant maintenant les équations (17.39). Il vient :

$$\sum_{i=1}^{n} \lambda_i = 0, \quad \sum_{i=1}^{n} \lambda_i \widehat{x}_i = 0. \tag{17.51}$$

On a $\sum \lambda_i = \sum w_i z_i$, d'où on tire l'expression de $\widehat{b}$:

$$\widehat{b} = \frac{\sum_{i=1}^{n} \widehat{w_i} y_i - \widehat{a} \sum_{i=1}^{n} \widehat{w_i} x_i}{\sum_{i=1}^{n} \widehat{w_i}}, \tag{17.52}$$

où $\widehat{w}_i = w_i(\widehat{a})$. Cette dernière équation peut également s'écrire :

$$\widehat{b} = \bar{y} - \widehat{a}\bar{x}, \tag{17.53}$$

avec :

$$\bar{y} = \frac{\sum_i \widehat{w}_i y_i}{\sum_i \widehat{w}_i}, \quad \bar{x} = \frac{\sum_i \widehat{w}_i x_i}{\sum_i \widehat{w}_i}, \tag{17.54}$$

ce qui montre que la droite cherchée passe par le centre de gravité du nuage de points, chaque point étant affecté du poids $\widehat{w}_i$ à déterminer.

L'expression (17.49) définissant les poids montre qu'ils sont toujours positifs et que par conséquent le centre de gravité est dans l'enveloppe convexe du nuage de points. Il reste à évaluer $\widehat{a}$ que l'on tire de $\sum \lambda_i \widehat{x}_i = 0$, et qui montre que $\widehat{a}$ est solution du pseudo-polynôme $Q(a)$ suivant :

$$\begin{aligned} Q(a) = a^3 & \sum_{i=1}^{n} \sigma_{xi}^2 w_i^2 x_i^2 \\ & + a^2[-2 \sum_{i=1}^{n} \sigma_{xi}^2 w_i^2 x_i (y_i - \widehat{b})] \\ & + a[\sum_{i=1}^{n} \sigma_{xi}^2 w_i^2 (y_i - \widehat{b})^2 - \sum_{i=1}^{n} w_i x_i^2] \\ & + \sum_{i=1}^{n} w_i x_i (y_i - \widehat{b}), \end{aligned} \tag{17.55}$$

et en posant $u_i = x_i - \bar{x}$, et $v_i = y_i - \bar{y}$, il vient :

$$\begin{aligned} Q(a) = a^3 & \sum_{i=1}^{n} \sigma_{xi}^2 w_i^2 u_i^2 \\ & - 2a^2 \sum_{i=1}^{n} \sigma_{xi}^2 w_i^2 u_i v_i \\ & + a[\sum_{i=1}^{n} \sigma_{xi}^2 w_i^2 v_i^2 - \sum_{i=1}^{n} w_i u_i^2] \\ & + \sum_{i=1}^{n} w_i u_i v_i \ . \end{aligned} \tag{17.56}$$

On obtient aisément la solution $Q(\widehat{a}) = 0$ par itération.

Annexe A

Fonctions spéciales.

A.1 Fonctions eulériennes.

Le domaine que nous étudions fait souvent appel à deux fonctions continûment et indéfiniment différentiables sur leurs domaines de définition respectifs : ce sont les fonctions eulériennes B et Γ. Nous ne considèrerons ici que les fonctions eulériennes définies sur $\mathbb{R}^+$; dans le cas général, on les définit sur le plan complexe, les pôles $-1, -2, -3 \cdots$ exceptés.

A.1.1 Fonction eulérienne de première espèce.

On appelle fonction eulérienne de première espèce la fonction B, définie pour tout couple x, y de $\Omega =]0, \infty[\times]0, \infty[$ par :

$$B(x, y) = \int_0^1 t^{x-1}(1-t)^{y-1}\, dt\ . \tag{A.1}$$

On appelle plus couramment « fonction bêta » la fonction B.

Propriétés de la fonction bêta.

La fonction B est symétrique $B(x, y) = B(y, x)$.

A.1.2 Fonction eulérienne de deuxième espèce.

On appelle fonction eulérienne de deuxième espèce la fonction Γ, définie sur $\Omega =]0, \infty[$ par :

$$\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}\, dt\ . \tag{A.2}$$

On appelle plus couramment « fonction gamma» la fonction Γ. La figure A.1 représente le logarithme de la fonction Γ pour $0 < x \leq 5$.

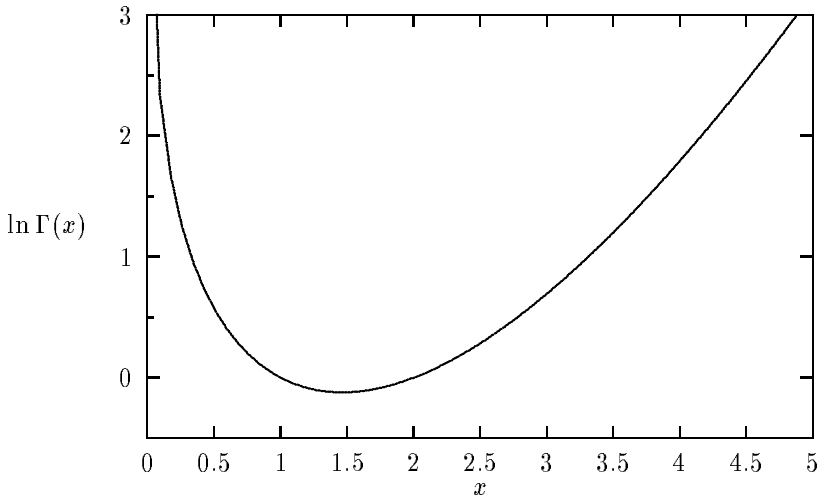


FIG. A.1: *Logarithme de la fonction* Γ.

Propriétés de la fonction gamma.

Formule des compléments. Pour tout réel $0 < x < 1$, on a :

$$\Gamma(1-x)\Gamma(x) = \frac{\pi}{\sin \pi x}, \quad 0 < x < 1\,, \tag{A.3a}$$

$$\Gamma(1-x)\Gamma(1+x) = \frac{\pi x}{\sin \pi x}, \quad 0 < x < 1\,. \tag{A.3b}$$

Ces relations permettent de connaître la fonction Γ sur $0 < x < 0.5$ quand on la connait sur $0.5 < x < 1$, et sur $0 < x < 1$ quand on la connaît sur $1 < x < 2$.

Formule de récurrence. Pour tout x réel positif, on a :

$$\Gamma(x+1) = x\Gamma(x), \quad x > 0\,. \tag{A.4}$$

Entiers et demi-entiers. Pour tout entier $n \geq 0$, on a :

$$\Gamma(n+1) = n!, \quad \Gamma(n+\tfrac{1}{2}) = \frac{(2n)!}{2^{2n}n!}\sqrt{\pi} = \frac{1}{2}\frac{3}{2}\cdots\frac{2n-1}{2}\sqrt{\pi}\,. \tag{A.5}$$

En particulier pour $n = 0$: $\Gamma(1) = 1$, et $\Gamma(\frac{1}{2}) = \sqrt{\pi}$.

Limites. Pour $x \to \infty$ on a :

$$\lim_{x\to\infty} \frac{\Gamma(x+\frac{1}{2})}{\Gamma(x)} = \sqrt{x} \tag{A.6}$$

Relation entre la fonction bêta et la fonction gamma.

Pour tout couple de réels positifs $x > 0, y > 0$, on a :

$$B(x,y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}\,. \tag{A.7}$$

A.2 Fonctions eulériennes incomplètes.

A.2.1 Fonction bêta incomplète.

On appelle fonction bêta *incomplète,* la fonction B_x définie pour $0 \leq x \leq 1$ et $a, b \in \mathbb{R}^+$:

$$B_x(a,b) = \int_0^x t^{a-1}(1-t)^{b-1}\,dt, \tag{A.8}$$

et fonction bêta incomplète *normalisée*, la fonction I_x :

$$I_x(a,b) = \frac{B_x(a,b)}{B(a,b)} \ . \tag{A.9}$$

Propriétés de la fonction bêta incomplète.

$$I_0(a,b) = 0, \quad I_1(a,b) = 1 \ . \tag{A.10}$$

Symétrie.

$$I_x(a,b) = 1 - I_{1-x}(b,a) \ . \tag{A.11}$$

Récurrence. On a la relation de récurrence suivante :

$$I_x(a,b) = xI_x(a-1,b) + (1-x)I_x(a,b-1) \ . \tag{A.12}$$

Relation avec la loi binomiale. On a les relations suivantes :

$$\sum_{k=0}^{r} C_n^k p^k (1-p)^{n-k} = 1 - I_p(r+1, n-r)\,, \tag{A.13}$$

$$\sum_{k=r}^{n} C_n^k p^k (1-p)^{n-k} = I_p(r, n-r+1) \ . \tag{A.14}$$

Ces relations permettent de calculer la fonction de répartition de la loi binomiale.

A.2.2 Fonction gamma incomplète.

On appelle fonction gamma *incomplète*, la fonction γ, définie pour $a, x \in \mathbb{R}^+$:

$$\gamma(a,x) = \int_0^x t^{a-1}e^{-t}\,dt, \tag{A.15}$$

et fonction gamma incomplète normalisée, la fonction P :

$$P(a,x) = \frac{\gamma(a,x)}{\Gamma(a)} \ . \tag{A.16}$$

Propriétés de la fonction gamma incomplète.

Relation avec la loi de Poisson. On a la relation suivante :

$$\sum_{k=0}^{r} \frac{\mu^k}{k!} e^{-\mu} = 1 - P(r+1, \mu) \; . \tag{A.17}$$

Cette relation permet de calculer la fonction de répartition de la loi de Poisson.

A.3 Fonction hypergéométrique.

La fonction hypergéométrique $F(\alpha, \beta, \gamma; z)$ est définie par la série suivante :

$$F(\alpha, \beta, \gamma\,; z) = 1 + \frac{\alpha\beta}{\gamma}\frac{z}{1!} + \frac{\alpha(\alpha+1)\beta(\beta+1)}{\gamma(\gamma+1)}\frac{z^2}{2!} + \cdots$$
$$\cdots + \frac{\alpha(\alpha+1)\ldots(\alpha+j-1)\beta(\beta+1)\ldots(\beta+j-1)}{\gamma(\gamma+1)\ldots(\gamma+j-1)}\frac{z^j}{j!} + \cdots \; , \tag{A.18}$$

où $\alpha, \beta, \gamma \in \mathbb{C}$, sauf un certain domaine précisé ci-dessous.

A.3.1 Domaine de définition.

- Si α et β ne sont pas des entiers négatifs ou nuls, mais si γ est un entier négatif ou nul, alors F n'est pas définie.

- Si α ou $\beta = -m$, et $\gamma = -n$ $(m, n \in \mathbb{N})$, alors F n'est pas définie si $-m < -n$. Si α et β sont des entiers négatifs ou nuls, on doit considérer $-m$ comme étant égal au plus grand des deux.

Convergence.

Dans le domaine des α, β, γ où F est définie, la série (A.18) converge dans le disque ouvert $|z| < 1$. Sur le cercle unité $|z| = 1$, on a les 3 cas suivants :

1. $\Re(\alpha + \beta - \gamma) < 0$: la série converge absolument sur tout le cercle unité.

2. $0 \leq \Re(\alpha+\beta-\gamma) < 1$: la série converge sur le cercle unité sauf pour $z = 1$.

3. $1 \leq \Re(\alpha + \beta - \gamma)$: la série diverge sur tout le cercle unité.

Dans le domaine des α, β, γ où F est définie et si α ou β sont des entiers négatifs ou nuls, la série (A.18) est finie et F est un polynôme en z défini sur tout le plan complexe.

A.3.2 Propriétés de la fonction hypergéométrique.

Pour $\alpha = \beta = \gamma = 1$ la série (A.18) devient une progression géométrique de raison z.

Valeurs particulières.

$$F(1,1,\tfrac{3}{2},\tfrac{1}{2}) = \frac{\pi}{2}\,. \tag{A.19}$$

$$F(\alpha,\beta,\gamma,1) = \frac{\Gamma(\gamma)\Gamma(\gamma-\alpha-\beta)}{\Gamma(\gamma-\alpha)\Gamma(\gamma-\beta)}, \quad \Re(\alpha+\beta-\gamma) < 0\,. \tag{A.20}$$

Equation différentielle. La fonction hypergéométrique $F(\alpha,\beta,\gamma,z)$ est une solution u_1 de l'équation différentielle suivante :

$$z(1-z)\frac{d^2u}{dz^2} + (\gamma - (\alpha+\beta+1)z)\frac{du}{dz} - \alpha\beta u = 0\,; \tag{A.21}$$

l'autre solution u_2 s'exprime, en général, aussi à l'aide d'une fonction hypergéométrique.

A.3.3 Fonction hypergéométrique généralisée.

L'expression de la fonction hypergéométrique généralisée ${}_rF_s$ est donnée par la série suivante :

$${}_rF_s(\alpha_1,\ldots,\alpha_r,\,\gamma_1,\ldots,\gamma_s\,;x) = \frac{\Gamma(\gamma_1)\ldots\Gamma(\gamma_s)}{\Gamma(\alpha_1)\ldots\Gamma(\alpha_r)}\sum_{n=1}^{\infty}\frac{\Gamma(\alpha_1+n)\ldots\Gamma(\alpha_r+n)}{\Gamma(\gamma_1+n)\ldots\Gamma(\gamma_s+n)}\frac{x^n}{n!}\,. \tag{A.22}$$

Pour $r=2$ et $s=1$ on obtient la fonction hypergéométrique introduite ci-dessus :

$${}_2F_1(\alpha_1,\alpha_2,\gamma\,;x) = F(\alpha_1,\alpha_2,\gamma\,;x) \tag{A.23}$$

A.3.4 Fonction hypergéométrique confluente.

Pour les valeurs $r=s=1$ de ${}_rF_s$ on obtient la fonction hypergéométrique confluente (ou dégénérée) ${}_1F_1$:

$${}_1F_1(\alpha,\gamma\,;x) = 1 + \frac{\alpha}{\gamma}\frac{x}{1!} + \frac{\alpha(\alpha+1)}{\gamma(\gamma+1)}\frac{x^2}{2!} + \cdots + \cdots + \frac{\alpha(\alpha+1)\ldots(\alpha+j-1)}{\gamma(\gamma+1)\ldots(\gamma+j-1)}\frac{x^j}{j!} + \cdots\,. \tag{A.24}$$

Les valeurs positives et négatives de cette fonction sont reliées entre-elles par la formule :

$${}_1F_1(\alpha,\gamma\,;x) = e^x\,{}_1F_1(\gamma-\alpha,\gamma\,;-x)\,. \tag{A.25}$$

Pour les grandes valeurs négatives de x on a le dévelppement asymptotique suivant :

$${}_1F_1(\alpha,\gamma\,;-x) \sim \frac{\Gamma(\gamma)}{\Gamma(\gamma-\alpha)}\frac{1}{x^\alpha}\Big[1 + \frac{\alpha(\alpha-\gamma+1)}{x} + \frac{\alpha(\alpha-\gamma+1)(\alpha-\gamma+2)}{2x^2} + \cdots\Big]\,. \tag{A.26}$$

Si les paramètres α et β sont entiers ou demi-entiers alors ${}_1F_1$ peut s'exprimer à l'aide de fonctions usuelles. Ainsi, par exemple :

$$ {}_1F_1(\tfrac{1}{2}, 1\,; -x) = e^{\frac{1}{2}x} I_0(\tfrac{1}{2}x)\,, \tag{A.27} $$

$$ {}_1F_1(-\tfrac{1}{2}, \tfrac{1}{2}, -\tfrac{1}{2}x^2) = e^{-\frac{1}{2}x^2} + x\sqrt{\frac{\pi}{2}}[2\Phi(x) - 1]\,, \tag{A.28} $$

où I_0 est la fonction de Bessel modifiée d'ordre 0 et Φ est la fonction de Laplace (fonction de répartition d'une variable aléatoire normale réduite).

A.4 Aspects numériques.

Les programmes cités ci-dessous peuvent être trouvés dans l'ouvrage *"Numerical Recipes,"* de Press *et al.* (1986) [60]. On y trouvera également les détails des méthodes numériques utilisées.

A.4.1 Fonction gamma.

La fonction Γ devient très vite plus grande que le plus grand réel représentable par un ordinateur, aussi préfère-t-on calculer le logarithme de cette fonction. Dans la pratique, la fonction Γ apparaît dans des rapports de grands nombres et ce rapport est souvent de l'ordre de l'unité. C'est donc bien la fonction $\ln\Gamma$ qui est vraiment utile.

Le programme `GAMMLN(X)` calcule la fonction $\ln\Gamma(x)$. Sa précision est de l'ordre de $\epsilon = 2 \times 10^{-10}$. Elle est meilleure pour $x > 1$ que pour $0 < x < 1$. Dans ce dernier domaine on s'aidera de la formule (A.3b) et on calculera :

$$ \ln\Gamma(x) = \ln\frac{\pi(1-x)}{\sin\pi(1-x)} - \ln\Gamma(2-x)\,. $$

A.4.2 Fonction bêta.

On calcule facilement la fonction B à l'aide du programme précédent et de la formule (A.7). Le programme `BETA(X,Y)` calcule cette fonction bêta : $B(x, y)$.

A.4.3 Fonction gamma incomplète.

Le programme `GAMMP(A,X)` calcule la fonction gamma incomplète normalisée $P(a, x)$.

A.4.4 Fonction bêta incomplète.

Le programme `BETAI(A,B,X)` calcule la fonction bêta incomplète normalisée $I_x(a, b)$.

Annexe B

Outils mathématiques.

B.1 Matrices.

Afin d'alléger l'exposé nous ne considérons que les matrices à éléments réels. Une matrice $\boldsymbol{A}$ à n lignes et m colonnes est un élément de $\mathbb{R}^{n,m}$. On note $\boldsymbol{x}$ un vecteur colonne et $\boldsymbol{x}^t$ un vecteur ligne. Un vecteur *unitaire* $\boldsymbol{u}$ est un vecteur tel que $\boldsymbol{u}^t\boldsymbol{u} = 1$.

B.1.1 Matrices définies positives.

Définition B.1. Une matrice carrée $\boldsymbol{M} \in \mathbb{R}^{n,n}$ est dite *définie positive* si, et seulement si, $\forall \boldsymbol{x} \neq \boldsymbol{0} \in \mathbb{R}^n$, la forme quadratique $\boldsymbol{x}^t\boldsymbol{M}\boldsymbol{x}$ est positive. On note $\boldsymbol{M} > 0$ pour indiquer que $\boldsymbol{M}$ est définie positive, on a alors :

$$\boldsymbol{M} > 0 \iff \forall \boldsymbol{x} \in \mathbb{R}^n, \quad \boldsymbol{x}^t\boldsymbol{M}\boldsymbol{x} = \sum_{i,j=1}^{n} M_{ij}x_ix_j > 0\,.$$

Dans cette définition, la seule restriction sur $\boldsymbol{M}$ est qu'elle soit carrée. Il est cependant possible de ne considérer que les matrices symétriques car l'expression $\boldsymbol{x}^t\boldsymbol{M}\boldsymbol{x}$ est invariante si l'on ajoute à $\boldsymbol{M}$ une matrice antisymétrique. En effet, si $\boldsymbol{B}^t = -\boldsymbol{B}$ on a : $\boldsymbol{x}^t\boldsymbol{B}\boldsymbol{x} = (\boldsymbol{x}^t\boldsymbol{B}\boldsymbol{x})^t = \boldsymbol{x}^t\boldsymbol{B}^t\boldsymbol{x} = -\boldsymbol{x}^t\boldsymbol{B}\boldsymbol{x} = 0$. On donne ci-dessous un ensemble de conditions nécessaires et suffisantes pour qu'une matrice symétrique soit définie positive.

Théorème B.1. *Pour qu'une matrice $\boldsymbol{A}$ réelle et symétrique soit définie positive, il faut et il suffit qu'une des conditions suivantes soit satisfaite.*

0. $\forall \boldsymbol{x} \neq \boldsymbol{0} \in \mathbb{R}^n, \quad \boldsymbol{x}^t\boldsymbol{A}\boldsymbol{x} > 0.$

1. *$\boldsymbol{A}$ possède des valeurs propres positives :* $\lambda_1 \geq \lambda_2 \cdots \geq \lambda_n > 0.$

2. *Les éléments diagonaux a_{ii} de $\boldsymbol{A}$ sont positifs et les éléments en dehors de la diagonale $a_{i\neq j}$ satisfont l'inégalité :* $a_{ij}^2 < a_{ii}a_{jj}$.

3. *Les déterminants principaux de $\boldsymbol{A}$ sont positifs. C'est-à-dire :*

$$a_{11} > 0, \quad \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} > 0, \quad \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} > 0, \quad \ldots \quad \det \boldsymbol{A} > 0\,.$$

Définition B.2. Une matrice carrée $\boldsymbol{M}$ est dite *semi-définie positive* (où définie non-négative) si, et seulement si, $\forall \boldsymbol{x}$ la forme quadratique $\boldsymbol{x}^t \boldsymbol{M} \boldsymbol{x}$ est positive ou nulle et il existe au moins un vecteur $\boldsymbol{z} \neq \boldsymbol{0}$ tel que $\boldsymbol{z}^t \boldsymbol{M} \boldsymbol{z} = 0$. On note $\boldsymbol{M} \geq 0$ une matrice semi-définie positive.

De la même façon que pour les matrices définies positives on peut, pour les matrices semi-définies positives, ne s'intéresser qu'à la classe des matrices symétriques. Il existe de très nombreux théorèmes concernant les matrices symétriques définies positives et semi-définies positives (on consultera, par exemple, l'appendice A de Rao & Toutenburg [61]).

B.1.2 Matrices projectives.

Définition B.3. Une matrice carrée $\boldsymbol{A}$ est dite matrice projective si elle est idempotente, c'est-à-dire si : $\boldsymbol{A}^2 \stackrel{\text{def}}{=} \boldsymbol{A}\boldsymbol{A} = \boldsymbol{A}$.

Une matrice projective symétrique est apellée *projecteur orthogonal*, dans tous les autres cas c'est un *projecteur oblique*.

Théorème B.2. *Soit $\boldsymbol{A} \in \mathbb{R}^{n,n}$ une matrice projective de rang* $\operatorname{rg} \boldsymbol{A} = r \leq n$. *On a :*

0. $\boldsymbol{A}^2 = \boldsymbol{A}$,

1. Les valeurs propres de $\boldsymbol{A}$ valent 0 ou 1,

2. La trace de $\boldsymbol{A}$ est égale à son rang, $\operatorname{trace} \boldsymbol{A} = \operatorname{rg} \boldsymbol{A}$,

3. Si $\boldsymbol{A}$ est de rang n, alors $\boldsymbol{A} = \boldsymbol{I}$,

4. $\boldsymbol{B} = \boldsymbol{I} - \boldsymbol{A}$ est aussi une matrice projective, on a $\boldsymbol{A}\boldsymbol{B} = \boldsymbol{B}\boldsymbol{A} = \boldsymbol{0}$.

La matrice $\boldsymbol{I} - \boldsymbol{u}\boldsymbol{u}^t$, où $\boldsymbol{u}$ est un vecteur unitaire, est un projecteur orthogonal (dans la direction du vecteur $\boldsymbol{u}$).

B.1.3 Inverses généralisées.

Définition B.4. Soit une matrice rectangulaire $\boldsymbol{A} \in \mathbb{R}^{m,n}$. On appelle inverse généralisée la matrice rectangulaire $\boldsymbol{A}^- \in \mathbb{R}^{n,m}$, telle que :

$$\boldsymbol{A}\boldsymbol{A}^-\boldsymbol{A} = \boldsymbol{A}\,.$$

Une matrice $\boldsymbol{A}$ quelconque admet toujours une inverse généralisée, mais elle n'est pas nécessairement unique.

En multipliant l'expression précédente à gauche et à droite par $\boldsymbol{A}^-$, on montre que les matrices carrées $\boldsymbol{A}^-\boldsymbol{A}$ et $\boldsymbol{A}\boldsymbol{A}^-$ sont des matrices projectives.

Théorème B.3. *(Moore-Penrose) Etant donné une matrice $\boldsymbol{A}$ quelconque, il existe une et une seule matrice $\boldsymbol{A}^{(-1)}$, appelée matrice pseudo-inverse de $\boldsymbol{A}$, vérifiant les conditions suivantes :*

$$\boldsymbol{A}\boldsymbol{A}^{(-1)}\boldsymbol{A} = \boldsymbol{A} \qquad \boldsymbol{A}^{(-1)}\boldsymbol{A}\boldsymbol{A}^{(-1)} = \boldsymbol{A}^{(-1)}$$
$$(\boldsymbol{A}^{(-1)}\boldsymbol{A})^t = \boldsymbol{A}^{(-1)}\boldsymbol{A} \qquad (\boldsymbol{A}\boldsymbol{A}^{(-1)})^t = \boldsymbol{A}\boldsymbol{A}^{(-1)}$$

B.2 Eléments de topologie.

B.2.1 Espaces topologiques.

On dit qu'un ensemble E est un *espace topologique* si on a pu y définir la notion de partie ouverte. Une partie ouverte, ou plus simplement un *ouvert*, est un élément d'une famille $\mathcal{O}$ de parties de E possédant les propriétés suivantes :

O1. $E \in \mathcal{O}$, $\varnothing \in \mathcal{O}$, (*$E$ et le vide sont ouverts*),

O2. $A_i \in \mathcal{O} \Rightarrow \bigcap_{i=1}^{n} A_i \in \mathcal{O}$, (*une intersection finie d'ouverts est ouverte*),

O3. $A_x \in \mathcal{O} \Rightarrow \bigcup_x A_x \in \mathcal{O}$, (*une réunion quelconque d'ouverts est ouverte*).

Par « *réunion quelconque* » on entend toute réunion finie ou infinie dénombrable ou non. On appelle *topologie* une famille d'ouverts, et *partie fermée* (ou fermé) toute partie de E dont le complémentaire est ouvert. Il résulte de cette définition que E et $\varnothing$ sont à la fois ouverts et fermés.

B.2.2 Espaces métriques.

Un *espace métrique* est un ensemble quelconque E muni d'une *distance* entre un couple (x, y) de ses éléments. Une distance est une application d de $E \times E$ dans $\mathbb{R}$ qui est d'abord un *écart*, c'est-à-dire $\forall x, y, z \in E$:

1. $d(x, y) = d(y, x)$ (*symétrie*),
2. $d(x, y) \leq d(x, z) + d(z, y)$ (*inégalité triangulaire,*)
3. $d(x, y) \geq 0$ (*non négativité*).

Un écart peut être nul même si $x \neq y$. Pour que d soit une distance la condition 3 précédente doit être remplacée par :

3. $d(x, y) = 0 \Longrightarrow x = y$ (*non dégénérescence*).

La non négativité devient alors une conséquence de la définition.

On appelle *boule ouverte* de centre $O \in E$ et de rayon $r > 0$, la partie de E suivante : $B_O(r) = \{x \,|\, d(O, x) < r\}$. Une boule ouverte est un ouvert, tout espace métrique est donc un espace topologique.

B.3 Structures algébriques.

B.3.1 Espaces vectoriels.

Un *espace vectoriel* E sur un corps K ($K = \mathbb{R}$ ou $K = \mathbb{C}$) est un ensemble pour lequel existe une addition entre éléments de E et une mutiplication d'un élément du corps par un élément de E. Ces opérations doivent satisfaires les axiomes suivants :

1. Il faut qu'elles soient *stables*, c'est-à-dire, $\forall \vec{x}, \vec{y} \in E, \forall \alpha \in K$:

$$\vec{x} + \vec{y} \in E \quad \text{et} \quad \alpha\vec{x} \in E\,.$$

2. L'ensemble E doit être un *groupe abélien* pour l'addition. C'est-à-dire $\forall \vec{x}, \vec{y}, \vec{z} \in E$:

$$\vec{x} + \vec{y} = \vec{y} + \vec{x}, \quad (\vec{x} + \vec{y}) + \vec{z} = \vec{x} + (\vec{y} + \vec{z}),$$

et il doit exister un élément neutre $\vec{0}$ et un inverse $\vec{x}'$ appartenant à E, tels que :

$$\vec{x} + \vec{0} = \vec{0}, \quad \vec{x} + \vec{x}' = \vec{0}\,.$$

3. La multiplication par un élément du corps doit posséder les propriétés suivantes, $\forall \alpha, \beta \in K$, $\forall \vec{x}, \vec{y} \in E$:

$$(\alpha + \beta)\vec{x} = \alpha\vec{x} + \beta\vec{x}, \quad \alpha(\vec{x} + \vec{y}) = \alpha\vec{x} + \alpha\vec{y},$$
$$\alpha(\beta\vec{x}) = (\alpha\beta)\vec{x}, \quad 1\vec{x} = \vec{x}\,.$$

Les éléments d'un espace vectoriel s'appellent des *vecteurs*, on les note habituellement par des lettres minuscules de l'alphabet latin : $\vec{x}, \vec{y}, \cdots \in E$. Les éléments du corps sont des *nombres*, on les note habituellement par des lettres minuscules de l'alphabet grec : $\alpha, \beta, \cdots \in K$. Le corps lui-même est appelé *espace numérique*.

Enveloppe linéaire. Soit M un sous-ensemble non vide de E ($M \subseteq E$). On appelle *enveloppe linéaire* de M l'ensemble, $\mathrm{span}(M)$, de toutes les combinaisons linéaires finies des vecteurs de M :

$$\forall \alpha_i \in K, \quad \mathrm{span}(M) \stackrel{\mathrm{def}}{=} \sum_i \alpha_i \vec{x}_i \quad \text{où} \quad \forall i\ \vec{x}_i \in M\,.$$

On pose $\mathrm{span}(\varnothing) = \varnothing$. L'ensemble M est un *sous-espace vectoriel* si, et seulement si, $\mathrm{span}(M) = M$ et $M \neq \varnothing$.

Indépendance linéaire. Les vecteurs $\vec{x}_1, \ldots, \vec{x}_n$ d'un espace vectoriels sont dits linéairement indépendants si, et seulement si, :

$$\forall \alpha_i \in K, \quad \alpha_1\vec{x}_1 + \cdots + \alpha_n\vec{x}_n = 0 \implies \forall i,\ \alpha_i = 0\,.$$

Ainsi il n'est pas possible d'exprimer linéairement le vecteur $\vec{x}_i$ à l'aide des autres vecteurs $\vec{x}_{j\neq i}$.

La *dimension* de E, $\dim(E)$, est le nombre maximum (s'il existe) de vecteurs de E linéairement indépendants. Si ce nombre existe E est dit de dimension finie, E est dit de dimension infinie si $\forall n \in \mathbb{N}$ il existe n vecteurs de E linéairement indépendants. Par définition, la dimension du sous-espace vectoriel réduit au seul élément nul est nulle ($\dim(\{\vec{0}\}) = 0$).

Bases d'un espace vectoriel. Si E est de dimension finie ($\dim E = n$), on appelle *base* de E un ensemble de n vecteurs de E linéairement indépendants. Les vecteurs de E peuvent se décomposer de façon *unique* sur une base $\{\vec{e}_1, \ldots, \vec{e}_n\}$ de E :

$$\forall \vec{x} \in E, \quad \vec{x} = \alpha_1\vec{e}_1 + \cdots + \alpha_n\vec{e}_n\,.$$

Les nombres $\alpha_1, \ldots, \alpha_n$ sont appelés les *composantes* du vecteur $\vec{x}$ par rapport à la base $\{\vec{e}_1, \ldots, \vec{e}_n\}$.

Changement de base. Soit une base $\Delta = \{\vec{e}_1, \ldots, \vec{e}_n\}$ d'un espace vectoriel de dimension n, supposons que l'on trouve une nouvelle base $\Delta' = \{\vec{e}'_1, \ldots, \vec{e}'_n\}$ par l'intermédiaire des n combinaisons linéaires suivantes :

$$\vec{e}'_i = \sum_{j=1}^{n} a_i^j \vec{e}_j \quad i, j = 1, \ldots, n\,.$$

Si les x^i désignent les composantes d'un vecteur $\vec{x}$ sur la base Δ et les x'^j sur la base Δ', il vient :

$$x^i = \sum_{j=1}^{n} a_j^i x'^j\,,$$

On range les a_j^i dans une matrice carrée $\boldsymbol{A}$ de la façon suivante :

$$\boldsymbol{A} = \begin{pmatrix} a_1^1 & a_2^1 & \cdots & a_n^1 \\ a_1^2 & a_2^2 & \cdots & a_n^2 \\ \vdots & \vdots & \ddots & \vdots \\ a_1^n & a_2^n & \cdots & a_n^n \end{pmatrix}$$

Avec cette convention, la colonne numéro i de $\boldsymbol{A}$ est formée des composantes du nouveau vecteur $\vec{e}'_i$ suivant l'ancienne base Δ. Cette matrice, dite de *changement de base* permet de calculer les *anciennes* composantes de $\vec{x}$ en fonction des *nouvelles* alors qu'elle est définie comme permettant de calculer la *nouvelle* base à partir de l'*ancienne*. Pour cette raison les composantes d'un vecteur définies par l'enveloppe linéaire de la base, sont dites *contravariantes*.

Pour connaître les nouvelles composantes en fonction des anciennes il faut que la matrice inverse de $\boldsymbol{A}$ existe, c'est-à-dire : $\det \boldsymbol{A} \neq 0$, alors les vecteurs $\vec{e}'_i$ forment effectivement une base de l'espace vectoriel. Si $\boldsymbol{B} = \boldsymbol{A}^{-1}$, on a :

$$x'^i = \sum_{j=1}^{n} b_j^i x^j, \quad \text{soit, en notation matricielle} \quad \boldsymbol{x}' = \boldsymbol{B}\boldsymbol{x}\,.$$

Suivant la convention tensorielle, on note des composantes contravariantes avec un indice supérieur. Cette convention n'est malheureusement pas respectée en statistique où l'on utilise plutôt la notation matricielle.

B.3.2 L'espace dual.

Formes linéaires. Une forme (ou fonctionnelle) f est une application d'un espace vectoriel E dans l'espace numérique $\mathbb{C}$ (ou $\mathbb{R}$). On note $f(\vec{x})$ le résultat de cette application. Dans le cas particulier des formes linéaires, définies ci-dessous, ce nombre est noté $\langle f, \vec{x}\rangle$. Une forme f est une *forme linéaire* si :

1. $\langle f, \vec{x} + \vec{y}\rangle = \langle f, \vec{x}\rangle + \langle f, \vec{y}\rangle$,
2. $\langle f, \lambda\vec{x}\rangle = \lambda\langle f, \vec{x}\rangle$.

L'addition entre formes linéaires et la multiplication par un nombre sont définies de la façon suivante : $\langle f + g, \vec{x}\rangle = \langle f, \vec{x}\rangle + \langle g, \vec{y}\rangle$, $\langle \lambda f, \vec{x}\rangle = \lambda\langle f, \vec{x}\rangle$. Les formes linéaires munies de ces deux opérations, forment un espace vectoriel appelé *espace dual* de E, on le note E^*.

Bases de l'espace dual. Si E est un espace vectoriel de dimension finie, l'espace dual E^* possède alors la même dimension que E : $\dim E^* = \dim E$. Si $\Delta = \{\vec{e}_1, \ldots, \vec{e}_n\}$ est une base de E, l'ensemble $\Delta^* = \{e_1^*, \ldots, e_n^*\}$, où $e_i^* = \langle e_i^*, \vec{e}_j \rangle = \delta_{ij}$ est une base de E^*.

Toutes les formes linéaires f sur E s'exprime alors linéairement en fonction de la base, on a :

$$\forall f \in E^*,\ f = \sum_{i=1}^{n} f_i e_i^* \quad \text{avec} \quad f_i = \langle f, \vec{e}_i \rangle .$$

La suite ordonnée $(f_1, \ldots, f_n)$ des composantes de f par rapport à la base Δ^* est un vecteur noté $\boldsymbol{f}$.

Changement de base. Soit $\boldsymbol{A}$ la matrice de changement de base qui fait passer de l'ancienne base Δ à la nouvelle base Δ'. On trouve les nouvelles composantes f_i' de f en fonction des anciennes f_j, de la façon suivante :

$$f_i' = \sum_{j=1}^{n} a_i^j f_j, \quad \text{soit} \quad \boldsymbol{f}'^t = \boldsymbol{f}^t \boldsymbol{A} .$$

Les composantes de f varient en fonction de la matrice de changement de base, comme la base elle-même, pour cette raison elles sont dites *covariantes*. Suivant la convention tensorielle on note des composantes covariantes avec un indice inférieur.

B.3.3 Espace vectoriels normés.

Un espace vectoriel E sur K est dit normé s'il possède une *norme*. Une norme est une application ν de E dans $\mathbb{R}$, notée $||\cdot||$ qui possède d'abord les propriétés d'une *semi-norme*, c'est-à-dire $\forall \lambda \in K$, $\forall \vec{x} \in E$:

1. $||\lambda \vec{x}|| = |\lambda| \cdot ||\vec{x}||$ (*homogénéité absolue*),
2. $||\vec{x} + \vec{y}|| \leq ||\vec{x}|| + ||\vec{y}||$ (*inégalité de convexité*).

Il découle immédiatement de ces premières propriétés : $||\vec{x}|| \geq 0$, $||\vec{0}|| = 0$, $||-\vec{x}|| = ||\vec{x}||$ et $\big|\, ||\vec{x}|| - ||\vec{y}|| \,\big| \leq ||\vec{x} \pm \vec{y}|| \leq ||\vec{x}|| + ||\vec{y}||$.

La semi-norme d'un vecteur peut être nulle sans que ce vecteur soit nul. Pour être une norme ν doit en outre posséder la propriété suivante :

3. $||\vec{x}|| = 0 \implies \vec{x} = \vec{0}$ (*non dégénérescence*).

Un espace normé est un espace métrique, en effet $d(\vec{x}, \vec{y}) = ||\vec{x} - \vec{y}||$ est une distance dans E. La réciproque n'est en général pas vraie. Pour qu'une distance d entre deux éléments d'un espace vectoriel soit engendrée par une norme il faut et il suffit qu'elle remplisse les deux conditions supplémentaires suivantes :

1. $d(\vec{x} + \vec{z}, \vec{y} + \vec{z}) = d(\vec{x}, \vec{y})$ (*invariance par translation*),
2. $d(\lambda \vec{x}, \lambda \vec{y}) = |\lambda|\, d(\vec{x}, \vec{y})$ (*homogénéité absolue*).

Dans ces conditions, il existe alors une et une seule norme telle que $d(\vec{x}, \vec{y}) = ||\vec{x} - \vec{y}||$.

B.3.4 Formes hermitiennes et produit scalaire.

Forme antilinéaire Une forme antilinéaire (à symétrie hermitienne) sur un espace vectoriel E construit sur un corps K est une application qui à un couple $(\vec{x}, \vec{y})$ de vecteurs de E fait correspondre un élément de K (un réel ou plus généralement un complexe). C'est une application de $E \times E$ vers K qui doit satisfaire les axiomes suivants, $\forall \lambda \in K$ et $\forall \vec{x}, \vec{y}, \vec{z} \in E$:

1. $(\vec{x}|\vec{y}) = \overline{(\vec{y}|\vec{x})}$ (*symétrie hermitienne*),
2. $(\vec{x}|\lambda\vec{y}) = \lambda(\vec{x}|\vec{y})$ (*homogénéité*),
3. $(\vec{x}|\vec{y}+\vec{z}) = (\vec{x}|\vec{y}) + (\vec{x}|\vec{z})$ (*linéarité à droite*).

La notation $\overline{z}$ désigne le complexe conjugé de z. Il vient des axiomes que $(\vec{x}|\vec{y})$ est anti-linéaire à gauche : $(\lambda\vec{x}|\vec{y}) = \overline{\lambda}(\vec{x}|\vec{y}), \quad (\vec{x}+\vec{y}|\vec{z}) = (\vec{x}|\vec{z}) + (\vec{y}|\vec{z})$.

Forme hermitienne. On appelle *forme hermitienne* la quantité $(\vec{x}|\vec{x})$ déduite de la forme antilinéaire. L'égalité :

$$(\vec{x}|\vec{y}) = \frac{1}{4}[(\vec{x}+\vec{y}|\vec{x}+\vec{y}) - (\vec{x}-\vec{y}|\vec{x}-\vec{y})] + \frac{i}{4}[(\vec{x}+i\vec{y}|\vec{x}+i\vec{y}) - (\vec{x}-i\vec{y}|\vec{x}-i\vec{y})]\,,$$

montre qu'il y a équivalence entre l'existence d'une forme antilinéaire à symétrie hermitienne et d'une forme hermitienne. On a : $(\vec{0}|\vec{0}) = 0$, mais la réciproque n'est pas nécessairement vraie.

Forme définie. Pour être une forme définie $(\vec{x}|\vec{x})$ doit en outre satisfaire l'axiome :

4. $\vec{x} \neq \vec{0} \Longrightarrow (\vec{x}|\vec{x}) \neq 0$, (*forme définie*),

On a alors l'équivalence suivante : $(\vec{x}|\vec{x}) = 0 \iff \vec{x} = \vec{0}$.

Théorème B.4. *(Cauchy-Schwarz.) Soient $\vec{x}$ et $\vec{y}$ deux vecteurs appartenant à un espace vectoriel E et $(\vec{x}|\vec{y})$ une forme définie. Alors :*

$$|(\vec{x}|\vec{y})|^2 \leq (\vec{x}|\vec{x})(\vec{y}|\vec{y})\,, \tag{B.1}$$

l'égalité n'ayant lieu que si, et seulement si, $\vec{x}$ et $\vec{y}$ sont linéairement dépendants.

Par conséquent : $(\vec{x}|\vec{x})(\vec{y}|\vec{y}) \geq 0$, ce qui implique qu'une forme définie est : soit définie positive, soit définie négative.

Produit scalaire. Un produit scalaire est une forme définie positive.

5. $\forall \vec{x} \in E \quad (\vec{x}|\vec{x}) \geq 0$.

Matrice de Gram. Soit $\Gamma = \{\vec{x}_1, \ldots, \vec{x}_k\}$ un ensemble de k vecteur de E. On appele « matrice de Gram » de Γ la matrice $\boldsymbol{G}$ dont les élements g_{ij} sont égaux aux produits scalaires des vecteurs de Γ. On a :

$$\boldsymbol{G} = \begin{pmatrix} (\vec{x}_1|\vec{x}_1) & \ldots & (\vec{x}_1|\vec{x}_k) \\ \vdots & \ddots & \vdots \\ (\vec{x}_k|\vec{x}_1) & \ldots & (\vec{x}_k|\vec{x}_k) \end{pmatrix} .$$

La matrice $\boldsymbol{G}$ est définie non-négative, on a $\det \boldsymbol{G} \geq 0$ et de plus :

Théorème B.5. *Pour qu'un ensemble de vecteurs soit composé de vecteurs linéairement indépendants, il faut et il suffit que sa matrice de Gram ne soit pas singulière.*

$$[\sum_{i=1}^{k} \alpha_i \vec{x}_i = 0 \Rightarrow \forall i,\ \alpha_i = 0\ ] \iff \det \boldsymbol{G} \neq 0 .$$

Si l'ensemble est réduit à seulement deux vecteurs $\vec{x}$ et $\vec{y}$, le théorème précédent exprime l'inégalité de Cauchy-Schwarz.

B.3.5 Espaces préhilbertien.

L'existence d'un produit scalaire confère à un espace vectoriel de dimension finie ou infinie une structure d'espace *préhilbertien* (un espace *hibertien* est un espace préhilbertien complet). Un espace préhilbertien est *normé* par l'intermédiaire de la norme :

$$||\vec{x}|| = (\vec{x}|\vec{x})^{\frac{1}{2}} . \tag{B.2}$$

La réciproque n'est, en général, pas vraie. On a cependant le théorème suivant :

Théorème B.6. *Un espace vectoriel normé E est un espace préhilbertien si, et seulement si, la norme satisfait l'égalité du parallélogramme :*

$$||\vec{x} + \vec{y}||^2 + ||\vec{x} - \vec{y}||^2 = 2(||\vec{x}||^2 + ||\vec{y}||^2) . \tag{B.3}$$

Orthogonalité. Par définition deux vecteurs $\vec{x}, \vec{y} \in E$ sont dit *orthogonaux* ($\vec{x} \perp \vec{y}$) si, et seulement si, leur produit scalaire est nul : $(\vec{x}|\vec{y}) = 0 \iff \vec{x} \perp \vec{y}$. Le seul vecteur orthogonal à tous les vecteurs de E est le vecteur nul : $\forall \vec{y} \in E,\quad (\vec{y}|\vec{x}) = 0 \iff \vec{x} = \vec{0}$.

Théorème B.7. *(Pythagore.) Soit E un espace vectoriel muni d'un produit scalaire, on a :*

$$\forall \vec{x}, \vec{y} \in E,\quad (\vec{x}|\vec{y}) = 0 \implies ||\vec{x} + \vec{y}||^2 = ||\vec{x}||^2 + ||\vec{y}||^2 ,$$

où $||\cdot||$ est la norme induite par le produit scalaire. La réciproque n'est vraie que dans les espaces réels où : $(\vec{x}|\vec{y}) \in \mathbb{R}$.

Systèmes orthonormés. Un ensemble U fini ou infini de vecteurs d'un espace préhilbertien E s'appelle un *système orthonormé* si :

$$\forall \vec{u}_i, \vec{u}_j \in U, (\vec{u}_i|\vec{u}_j) = \delta_{ij} ,$$

où δ_{ij} est le symbole de Kronecker ($\forall i, j \in \mathbb{N}$, $\delta_{ii} = 1, \delta_{i\neq j} = 0$).

On démontre grâce au *processus d'orthogonalisation de Gram-Schmidt* que dans un espace préhilbertien E il existe toujours un système orthonormé ayant une infinité dénombrable d'éléments. Dans le cas des espaces de dimension n finie, un système orthonormé de n éléments constitue une base.

Soient $\{\vec{v}_1, \ldots, \vec{v}_n\}$ n vecteurs linéairement indépendants de E, on trouve le premier vecteur $\vec{u}_1$ par :

$$\vec{u}_1 = \frac{\vec{v}_1}{||\vec{v}_1||} ,$$

le deuxième est trouvé par :

$$\vec{u}'_2 = \vec{v}_2 - (\vec{u}_1|\vec{v}_2)\vec{u}_1 , \quad \vec{u}_2 = \frac{\vec{u}'_2}{||\vec{u}'_2||} .$$

On répète le processus pour tous les $i \leq n$ en appliquant la formule :

$$\vec{u}'_i = \vec{v}_i - \sum_{j=1}^{i-1} (\vec{u}_j|\vec{v}_i)\vec{u}_j , \quad \vec{u}_i = \frac{\vec{u}'_i}{||\vec{u}'_i||} .$$

B.3.6 Espaces unitaires.

On réserve le nom d'espace unitaire aux espaces préhilbertiens de dimension finie. Dans un espace unitaire E, il est possible d'exprimer la forme générale prise par les formes antilinéaires par l'intermédiaire du *tenseur métrique*.

Tenseur métrique. Soit $\Delta = \{\vec{e}_1, \ldots, \vec{e}_n\}$ une base de E et $\vec{x}, \vec{y} \in E$. On a $\vec{x} = \sum_{i=1}^n x_i \vec{e}_i$, $\vec{y} = \sum_{j=1}^n y_j \vec{e}_j$ et il vient, $(\vec{x}|\vec{y}) = \sum_{i,j=1}^n \overline{x}_i y_j (\vec{e}_i|\vec{e}_j)$. On pose $g_{ij} = (\vec{e}_i|\vec{e}_j)$, les g_{ij} sont les éléments du tenseur métrique. La symétrie hermitienne impose : $g_{ji} = \overline{g}_{ij}$.

Réciproquement, si l'on se donne n^2 nombres g_{ij} tels que $g_{ji} = \overline{g}_{ij}$, la forme $(\vec{x}|\vec{y}) = \sum_{ij} g_{ij}\overline{x}_i y_j$, où les x_i et les y_i sont les composantes de $\vec{x}$ et $\vec{y}$ par rapport à une certaine base, est une forme antilinéaire.

Le théorème B.1 donne une série de conditions nécessaires et suffisantes pour que des g_{ij}, définissant une forme antilinéaire, définissent de plus un produit scalaire. Une de ces conditions est que $|g_{ij}|^2 < g_{ii}g_{jj}$.

Norme induite. Dans le cas d'une base orthonormée on a $g_{ij} = \delta_{ij}$, le produit scalaire s'écrit alors $(\vec{x}|\vec{y}) = \overline{x}_1 y_1 + \cdots + \overline{x}_n y_n$ et la norme induite :

$$||\vec{x}|| \stackrel{\text{def}}{=} (\vec{x}|\vec{x})^{\frac{1}{2}} = \left[|x_1|^2 + \cdots + |x_n|^2\right]^{\frac{1}{2}} .$$

On peut considérer l'ensemble des composantes de $\vec{x}$ par rapport à une base comme étant un vecteur. La formule ci-dessus montre que la norme induite par le produit scalaire est identique à la norme euclidienne du vecteur des composantes de $\vec{x}$ par rapport à une base orthonormée.

B.3.7 Espaces vectoriels arithmétiques.

On appelle *espace vectoriel arithmétique*, un espace vectoriel dont les vecteurs sont des suites ordonnées de n nombres. Cet espace est noté $\mathbb{R}^n$ ou $\mathbb{C}^n$ suivant que ces nombres sont des nombres réels ou des nombres complexes.

Par exemple, dans un espace vectoriel de dimension finie n, les composantes d'un vecteur par rapport à une base forment une suite ordonnée de n nombres et constituent un vecteur d'un espace arithmétique.

Normes. Dans un espace arithmétique toutes les quantités suivantes sont des normes :

$$\forall n \in \mathbb{N}^+,\ \left[|x_1|^n + \cdots + |x_n|^n\right]^{\frac{1}{n}} \stackrel{\text{def}}{=} ||\vec{x}||_n\,,$$
$$\sup_i |x_i| \stackrel{\text{def}}{=} ||\vec{x}||_\infty\,.$$

Produit scalaire subordonné. Une matrice symétrique définie positive $\boldsymbol{A}$ définit un produit scalaire $(\cdot|\cdot)_A$ appelé *produit scalaire subordonné à la matrice* $\boldsymbol{A}$. On a :

$$(\vec{x}|\vec{y})_A = \sum_{ij} a_{ij} x_i y_j = \boldsymbol{x}^t \boldsymbol{A} \boldsymbol{y}\,.$$

L'inégalité de Cauchy-Schwarz s'applique à ce produit scalaire, nous en donnons plusieurs formes équivalentes dans le théorème suivant.

Théorème B.8. *Si $\boldsymbol{x}$ et $\boldsymbol{y}$ sont des vecteurs d'un espace vectoriel arithmétique $\mathbb{R}^n$ et si $\boldsymbol{A} \in \mathbb{R}^{n,n}$ est une matrice symétrique définie positive, alors on a les propriétés suivantes :*

0. $(\boldsymbol{x}^t\boldsymbol{y})^2 \leq (\boldsymbol{x}^t\boldsymbol{x})(\boldsymbol{y}^t\boldsymbol{y})$,

1. $(\boldsymbol{x}^t\boldsymbol{A}\boldsymbol{y})^2 \leq (\boldsymbol{x}^t\boldsymbol{A}\boldsymbol{x})(\boldsymbol{y}^t\boldsymbol{A}\boldsymbol{y})$,

2. $(\boldsymbol{x}^t\boldsymbol{y})^2 \leq (\boldsymbol{x}^t\boldsymbol{A}\boldsymbol{x})(\boldsymbol{y}\boldsymbol{A}^{-1}\boldsymbol{y})$,

3. $\sup_{\boldsymbol{x}\neq\boldsymbol{0}} \dfrac{(\boldsymbol{x}^t\boldsymbol{y})^2}{\boldsymbol{x}^t\boldsymbol{A}\boldsymbol{x}} = \boldsymbol{y}^t\boldsymbol{A}^{-1}\boldsymbol{y}$,

4. $\sup_{\boldsymbol{x}^t\boldsymbol{A}\boldsymbol{x}=1} (\boldsymbol{x}^t\boldsymbol{y})^2 = \boldsymbol{y}^t\boldsymbol{A}^{-1}\boldsymbol{y}$.

B.4 Applications linéaires.

Une application $A : X \mapsto Y$, d'un espace vectoriel X dans un autre espace vectoriel Y, est une *application linéaire* si, et seulement si, elle possède les propriétés suivantes :

1. $A(\vec{x} + \vec{y}) = A\vec{x} + A\vec{y}$,
2. $A(\lambda\vec{x}) = \lambda A\vec{x}$.

Continuité. L'application A est continue en x_0 si $||A\vec{x} - A\vec{x}_0|| < \epsilon$ dès que $||\vec{x} - \vec{x}_0|| < \eta(\epsilon, \vec{x}_0)$. Si η ne dépend pas de $\vec{x}_0$, la continuité est uniforme.

Théorème B.9. *Soit $A : X \mapsto Y$ une application linéaire, alors les propositions suivantes sont équivalentes :*

1. *A est continue à l'origine,*
2. *$\exists m;\ ||A\vec{x}|| \leq m||\vec{x}||$, (application bornée),*
3. *A est continue uniformément.*

Définition B.5. Noyau. L'ensemble des solutions de l'équation $A\vec{z} = \vec{0}$ constitue un sous-espace vectoriel, on l'appelle *noyau* de l'application A et on le note $\ker A$.

$$\ker A \stackrel{\text{def}}{=} \{\vec{z} \,|\, A\vec{z} = \vec{0}\}\,.$$

Définition B.6. Image. L'ensemble des vecteurs $\vec{x}$ de la forme $\vec{y} = A\vec{x}$ est un sous-espace vectoriel, on le note $\text{ima}\, A$. C'est l'ensemble de tous les vecteurs de Y qui peuvent être « atteints » par A.

$$\vec{y} \in \text{ima}\, A \iff \exists \vec{x} \in Y;\ A\vec{x} = \vec{y}\,.$$

Définition B.7. Rang. On appelle *rang* de l'application linéaire A la dimension de $\text{ima}\, A$, (si celle est finie). On note $\text{rg}\, A$ le rang de A : $\text{rg}\, A \stackrel{\text{def}}{=} \dim \text{ima}\, A$.

Théorème B.10. Fredholm. *Pour que l'équation $A\vec{x} = \vec{b}$ admette une solution unique pour tout $\vec{b}$ de E, il faut et il suffit que l'équation homogène : $A\vec{z} = \vec{0}$, n'admette que la solution triviale $\vec{z} = \vec{0}$.*

B.4.1 Application adjointe.

Pour toute application linéaire continue, $A : X \mapsto Y$, il existe une application linéaire continue A^* telle que :

$$\langle f, Ax\rangle = \langle A^* f, x\rangle, \quad \forall f \in Y^*, x \in X\,.$$

L'application A^* est appelée application *adjointe* de A. On a $(\lambda A)^* = \lambda A^*$, $(A + B)^* = A^* + B^*$, $(AB)^* = B^* A^*$.

B.4.2 Espaces de dimensions finies.

Forme matricielle. Il est possible de représenter, à l'aide d'une matrice, toutes les applications linéaires d'un espace de dimension finie dans un autre espace de dimension finie.

Soit $A : X_n \mapsto Y_m$ une application linéaire d'un espace vectoriel de dimension n dans un espace vectoriel de dimension m. Soient $(x^1, \ldots, x^n)$ les composantes d'un vecteur $\vec{x}$ de X_n suivant une base Δ_X et $(y^1, \ldots, y^m)$ les composantes du vecteur $\vec{y} = A\vec{x}$ suivant une base Δ_Y de Y_m. Il vient, d'après la linéarité de A :

$$y^j = \sum_{i=1}^{n} a_i^j x^i, \quad j = 1, \ldots, m\,.$$

On range les a_i^j sous forme d'une matrice rectangulaire à m lignes et n colonnes, de la façon suivante :

$$\boldsymbol{A} = \begin{pmatrix} a_1^1 & \dots & a_n^1 \\ \dots\dots\dots\dots & & \\ a_1^m & \dots & a_n^m \end{pmatrix}$$

La colonne numéro i de $\boldsymbol{A}$ est égale aux composantes du vecteur $A\vec{e}_i$ sur la base Δ_Y de Y_m.

En notant $\boldsymbol{x}$ et $\boldsymbol{y}$ les vecteurs colonnes des composantes de $\vec{x}$ suivant Δ_X et de $\vec{y}$ suivant Δ_Y, on peut écrire l'équation $\vec{y} = A\vec{x}$ de façon équivalente, sous la forme: $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$. Ainsi une application linéaire quelconque est entièrement caratérisée par son effet sur une base de l'espace de départ.

La représentation matricielle de A dépend des bases choisies pour X_n et Y_m. Si $\boldsymbol{S}$ et $\boldsymbol{P}$ désignent les matrices de changement de base respectivement dans X_n et Y_m, on trouve la nouvelle représentation $\boldsymbol{A}'$ de A suivant les nouvelles bases, grâce à la formule suivante :

$$\boldsymbol{A}' = \boldsymbol{P}^{-1}\boldsymbol{A}\boldsymbol{S}\,, \quad \text{où} \quad \boldsymbol{x} = \boldsymbol{S}\boldsymbol{x}',\ \boldsymbol{y} = \boldsymbol{P}\boldsymbol{y}'\,.$$

Pour une application $A : X_n \mapsto Y_m$, on a les Propriétés suivantes :

1. Le rang de l'application est égal au rang de la matrice qui la représente :

$$\mathrm{rg}\, A = \mathrm{rg}\, \boldsymbol{A}\,.$$

2. La dimension de l'espace de départ est égale à la somme de la dimension de l'espace image (c'est-à-dire du rang de A) et de la dimension du noyau :

$$\dim \mathrm{ima}\, A + \dim \ker A = \dim X_n\,.$$

Annexe C

Eléments biographiques.

BAYES Thomas (Londres 27.12.1702 - Turnbridge Wells 17.04.1761), mathématicien britannique.

BERNOULLI Jakob I (Bâle 27.12.1654 - Bâle 16.08.1705), mathématicien suisse d'origine néerlandaise.

BERNSTEIN Sergueï Natanovitch (Odessa, Russie 05.03.1880 - Moscou URSS 26.10.1968), mathématicien et statisticien soviétique.

BESSEL Friedrich Wilhelm (Minden, auj. en Allemagne 22.07.1784 - Königsberg, auj. Kaliningrad en Russie 17.03.1846), astronome allemand.

BIENAYMÉ Irénée Jules (Paris 28.08.1796 - Paris 19.10.1878), statisticien français.

BORTKIEWICZ Ladislaus Josephowitch, von (St-Pétersbourg, Russie 07.08.-1868 - Berlin 15.07.1931), statisticien allemand d'origine polonaise.

CANTELLI Francesco Paolo (Palerme 20.12.1875 - Rome 21.07.1966), mathématicien italien.

CAUCHY Augustin Louis, Baron (Paris 21.08.1789 - Sceaux 23.05.1857), mathématicien et physicien français.

CRAMÉR Carl Harald (Stockholm 25.09.1893 - 1985), mathématicien suédois.

DARMOIS Georges (Eply 20.06.1888 - Paris 03.01.1960), mathématicien et physicien français.

DIRAC Paul Adrien Maurice (Bristol, Angleterre 08.08.1902 - Thallahasee, Floride 20.10.1984), physicien et mathématicien britannique.

ERLANG Agner Krarup (1878 - 1929), ingénieur suédois.

EUCLIDE (Alexandrie ? *ca*330 - *ca*260), mathématicien grec.

EULER Leonhard (Bâle 15.04.1707 - St-Pétersbourg 18.09.1783), mathématicien suisse.

FELLER William (Zagreb 07.14.1906 - New-York 14.01.1970), mathématicien américain d'origine croate.

FISHER Ronald Aylmer, Sir (Londres 17.02.1890 - Adelaïde, Australie 29.07.-1962) statisticien et généticien britannique.

FOURIER (Jean-Baptiste) Joseph, Baron (Auxerre 21.03.1768 - Paris 16.05.-1830), mathématicien et physicien français.

FRÉCHET (René) Maurice (Maligny, Yonne 02.09.1878 - Paris 1973), mathématicien français.

FREDHOLM Erik Ivar (Stockholm 07.04.1866 - Mörby 17.08.1927), mathématicien suédois.

GAUSS Karl Friedrich (Brunswick 30.04.1777 - Göttingen 23.02.1855), astronome, mathématicien et physicien allemand.

GOSSET William Sealy dit *"Student"* (Canterbury 13.06.1876 - Beaconsfield, Angleterre 16.10.1937), statisticien britannique.

GRAM Jòrgen Pedersen (Nastrup, auj. Haderslev, Danemark 27.06.1850 - Copenhague 29.04.1916), mathématicien danois.

HEAVISIDE Oliver (Londres 18.05.1850 - Paignton, Devon 03.02.1925), ingénieur électronicien, mathématicien et physicien britannique.

HELMERT Friedrich Robert (Freiberg, Saxe 31.07.1843 - Potsdam, Allemagne 15.06.1917), géodésien et astronome allemand.

HESSE Ludwig Otto (Königsberg auj. Kaliningrad, Russie 22.04.1811 - Munich 4.08.1874), mathématicien allemand.

HILBERT David (Königsberg auj. Kaliningrad, Russie 23.01.1862 - Göttingen 14.02.1943), mathématicien allemand.

HUYGENS Christiaan (La Haye 14.04.1629 - La Haye 08.07.1695), physicien, mathématicien et astronome néerlandais.

JACOBI Karl Gustav Jacob, von (Postdam 10.12.1804 - Berlin 18.02.1851), mathématicien allemand.

JENSEN Johan Ludwig William Voldemar (Nakskov, Danemark 08.05.1859 - Copenhague 05.03.1925), mathématicien danois.

KHINCHINE Aleksandr Yakovlevitch (Kondrovo, Kaluzhskaya Guberniya, Russie 19.07.1894 - Moscou 18.11.1959), mathématicien soviétique.

KOLMOGOROV Andreï Nikolaïevitch (Tambov, Russie 25.04.1903 - Moscou 20.10.1987), mathématicien soviétique.

KRONECKER Leopold (Liegnitz, auj. Legnica, Pologne 07.12.1823 - Berlin 29.12.1891), mathématicien allemand.

LAGRANGE Joseph Louis, comte de (Turin 25.01.1736 - Paris 10.04.1813), mathématicien et physicien théoricien français.

LAPLACE Pierre Simon, marquis de (Beaumont-en-Auge 23.03.1749 - Paris 05.03.1827), mathématicien, physicien et astronome français.

LEBESGUE Henri Léon (Beauvais 28.06.1875 - Paris 26.07.1941), mathématicien français.

LEVY Paul Pierre (Paris 15.09.1886 - Paris 15.12.1971), mathématicien français.

LIAPOUNOV Alexandre Mikhaïlovitch (Yaroslav, Russie 06.06.1857 - Odessa, URSS 03.11.1918), mathématicien et mécanicien russe.

LINDEBERG Jarl Waldemar (Helsinki 04.09.1876 - Helsinki 24.12.1932), mathématicien finlandais.

MARKOV Andreï Andreïevitch (Rjäsan, Russie 14.06.1856 - Pétrograd auj. St Pétersbourg 20.07.1922), mathématicien russe.

MAXWELL James Clerk (Edimbourg 13.06.1831 - Cambridge, Angleterre 05.-11.1879), physicien britannique.

MÉRÉ Antoine GOMBAUD, chevalier de (en Poitou *ca*1607 - Baussay, Poitou *ca*1685), Ecrivain français.

MOIVRE Abraham de (Vitry le françois 26.05.1667 - Londres 27.11.1754), mathématicien britannique d'origine française.

MOORE Eliakim Hastings (Marietta, Ohio 26.01.1862 - Chicago, Illinois 30.12.-1932), mathématicien américain.

NEWTON Isaac, Sir (Woolsthorpe, Lincolnshire, 25.12.1642 - Londres 20.03.-1727), Physicien, mathématicien et astronome anglais.

NEYMAN Jerzy (Bendery 16.04.1894 - Berkeley 5.08.1981), statisticien américain d'origine roumaine.

PEARSON Karl (Londres 27.03.1857 - Coldharbour, Angleterre 27.04.1936), biométricien britannique.

POISSON Siméon Denis (Pithiviers 21.06.1781 - Sceaux 25.04.1840), mathématicien et physicien français.

PÓLYA George (Budapest 13.12.1887 - Palo-Alto, Californie 07.09.1985), mathématicien américain d'origine hongroise.

PYTHAGORE (Samos *ca.*580 - Megapontum *ca.*500), philosophe et mathématicien grec.

QUETELET Lambert Adolphe Jacques (Gand 22.02.1796 - Bruxelles 1874), astronome et statisticien belge.

SCHWARZ (Karl) Hermann Amandus (Hermsdorf, Silésie 25.01.1843 - Berlin 30.11.1921), mathématicien allemand.

STIELTJES Thomas Jan (Zwolle, Overijssel, Pays-Bas 29.12.1856 - Toulouse 31.12.1894), mathématicien hollandais, naturalisé français.

STUDENT pseudonyme de *GOSSET W. S.*

TAYLOR Brook (Edmonton 18.08.1685 - Londres 29.12.1731), mathématicien britannique.

TCHÉBYCHEV Pafnouti Lvovich (Okatovo, Kalonga 16.05.1821 - St-Pétersbourg 8.12.1894), mathématicien russe.

Bibliographie

[1] Abel J.S., *A Bound on Mean-Square-Estimate Error*, IEEE Trans. Inform. Theory, 39 (1993), pp. 1675–1680.

[2] Abramowitz M. and Stegun I., *Handbook of Mathematical Functions*, National Bureau of Standards, 1970.

[3] Albert A., *Regression and the Moore-Penrose pseudoinverse*, Academic Press, New-York, 1972.

[4] Bayes T., *An essay towards solving a Problem in the Doctrine of Chances*, Phil. Trans., 53 (1763 (publié en 1764)), pp. 370–418. Reprint : Biometrika, 45 (1958), pp. 293–315.

[5] Beklémichev D., *Cours de géométrie analytique et d'algèbre linéaire*, Editions Mir, Moscou, 1988.

[6] Bellman R., *Introduction to Matrix Analysis*, Mc Graw-Hill, New-York, 2nd ed., 1970, p. 129.

[7] Bernoulli J., *Ars Conjectandi*, Thurnisiorum, Basel, 1713.

[8] Bhattacharyya A., *On some analogues of the amount of information and their use in statistical estimation*, Sankhyā, 8 (1946), pp. 1–14.

[9] ——, *idem*, Sankhyā, 8 (1947), pp. 201–211.

[10] ——, *idem*, Sankhyā, 8 (1948), p. 315.

[11] Borovkov A., *Statistique mathématique*, Editions Mir, Moscou, 1987.

[12] Bortkiewicz L. von, *Das Gesetz der kleinen Zahlen*, B.G. Teubner, Leipzig, 1898.

[13] Box G.E.P.and Muller M.E., *A note on the generation of random normal deviates*, Ann. Math. Statis., 29 (1958), pp. 610–611.

[14] Calot G., *Cours de calcul des probabilités*, Dunod, Paris, 2e ed., 1995.

[15] Cash W., Astrophys. J., 228 (1979), p. 939.

[16] Charles B. and Escoufier Y., *Probabilité et statistique*. Cours de l'Université des sciences et techniques du Languedoc, 1971.

[17] Cramér H., *Sur une propriété de la loi de Gauss*, C. R. Acad. Sci. Paris, 202 (1936), pp. 615–616.

[18] Dale A. I., *A History of Inverse Probability*, Springer-Verlag, New-York, 1995.

[19] Droesbeke J.-J., Fichet B., and Tassi Ph., eds., *Analyse statistique des durées de vie*, Economica, Paris, 1989.

[20] Feigelson E.D. and Nelson P.I., *Statistcal methods for astronomical data with upper limits. I. Univariate distribution*, Astrophys. J., 293 (1985), pp. 192–206.

[21] Fisher R.A., Biometrika, 30 (1915), p. 190.

[22] ——, *Frequency-distribution of the values of the correlation coefficient in samples from an indefinitely large population*, Biometrika, 10 (1915), p. 507.

[23] ——, *Statistical methods for research workers*, Oliver & Boyd, Edinburgh, 1925.

[24] Fisher R.A. and Tippett L.H.C., *Limiting form of the frequency-distribution of largest or smallest member of a sample*, Proc. Camb. Phil. Soc., 24 (1928), p. 180.

[25] Glazman I. and Liubitch V., *Analyse linéaire dans les espaces de dimensions finies*, Editions Mir, Moscou, 1972.

[26] Gnedenko B.V., *Sur la distribution limite du terme maximum d'une série aléatoire*, Ann. Math. (2), 44 (1943), p. 423.

[27] Gradshteyn I.S. and Ryzhik I.M., *Table of Integrals, Series, and Products*, Academic Press, New-York, 2nd ed., 1980.

[28] Hamming R.W., *On the distribution of numbers*, Bell. Syst. Tech. J., 49 (1970), pp. 1609–1625.

[29] Hardy G.H., Littlewood J.E., and Pólya G., *Inequalities*, Cambridge Mathematical Library, Cambridge, 2nd ed., 1951.

[30] Helmert F.R., Schämilch's für Math. und Physik, 20 (1875 a), p. 300.

[31] ——, Schämilch's für Math. und Physik, 21 (1875 b), p. 192.

[32] Hodge P.W., Astron. J., 88 (1983), p. 1323.

[33] Hotelling H., *New light on the correlation coefficient and its transform*, J. R. Statist. Soc. **B**, 15 (1953), p. 193.

[34] Jager O.C. de *et al.*, Astron. Astrophys., 170 (1986), p. 187.

[35] James F., *A review of pseudorandom generators*, Computer Phys. Comm., 60 (1990), pp. 329–344.

[36] Johnson N.L., Kotz S., and Balakrishnan N., *Continous Univariate Distributions*, vol. 1, Wiley, New-York, 2nd ed., 1994.

[37] ——, *Continous Univariate Distributions*, vol. 2, Wiley, New-York, 2nd ed., 1995.

[38] Johnson N.L., Kotz S., and Kemp A.W., *Univariate Discrete Distributions*, Wiley, New-York, 2nd ed., 1992.

[39] Kaplan E.L. and Meier P., *Nonparametric estimation from incomplete observations*, J. Am. Statist. Assoc., (1958), pp. 457–481.

[40] Kendall M. and Stuart A., *The advanced theory of statistics*, vol. 1, Ch. Griffin & Cie. Ltd., London & High Wycombe, 1977.

[41] ——, *The advanced theory of statistics*, vol. 2, Ch. Griffin & Cie. Ltd., London & High Wycombe, 1979.

[42] Khintchine A., *Über einen Satz der Wahrscheinlichkeitsrechnung*, Fundamenta Mathematiæ, 6 (1923), pp. 9–20.

[43] Kolmogorov A.N., *Grundbegriffe der Wahrscheinlichkeitsrechnung*, Springer, Berlin, 1933.

[44] Koroliouk V., *Aide-mémoire de théorie des probabilités et de statistique mathématique*, Editions Mir, Moscou, 1983.

[45] Kreijger R.G. and Neudecker H., *Exact linear restrictions on parameters in the general linear model with a singular covariance matrix*, J. Am. Stat. Assoc., 72 (1977), pp. 430–432.

[46] Laplace P.S., *Théorie analytique des probabilités*, V^{e} Courcier, Paris, 1812.

[47] Le Cam L., *The central limit theorem around 1935*, Statistical Science, 1 No.1 (1986), pp. 78–96.

[48] Lévine B., *Fondements Théoriques de la Radiotechnique Statistique*, vol. I, Editions Mir, Moscou, 1973.

[49] Loève M., *Probability theory*, Van Nostrand, Princeton, 3rd ed., 1963.

[50] Miller R.G., *The jackknife–a review*, Biometrika, 61 (1974), p. 1.

[51] Moivre A. de, *The doctrine of chance: or, A method of calculating the probability of events in play*, W. Pearson, London, 1718.

[52] Olkin I. and Pratt J.W., *Unbiased estimation of certain correlation coefficients*, Ann. Math. Statist., 29 (1958), p. 201.

[53] Papoulis A., *Probability, Random Variables, and Stochastic Processes*, Mc Graw-Hill, New-York, 2nd international student ed., 1984.

[54] Parzen E., *Modern Probability Theory and its Applications*, John Wiley & Sons, New-York, 1960.

[55] ——, *On estimation of a probability density function and mode*, Ann. Math. Statist., 33 (1962), pp. 1065–1076.

[56] Pascal B., *Œuvres complètes t. I*, Gallimard, Bibliothèque de la Plëïade, Paris, 1998.

[57] Pearson K., ed., *Tables of the incomplete beta-function*, Cambridge University Press, 1934.

[58] Poisson S.D., *Recherches sur la probabilité des jugements en matière criminelle et en matière civile*, Bachelier, Paris, 1837.

[59] Pólya G., *Ueber den zentralen grenzwertsatz der wahrscheinlichkeitsrechnung und das momentproblem*, Math. Z., 8 (1920), pp. 171–180.

[60] Press W., Flannery B., Teukolsky S., and Vetterling W., *Numerical Recipes*, Cambridge University Press, New-York, 1986.

[61] Rao C.R. and Toutenburg H., *Linear Models*, Springer, New York, 1995.

[62] Rényi A., *Calcul des probabilités*, Dunod, Paris, 1966.

[63] Rosenblatt M., *Remarks on some non parametric estimates of a density function*, Ann. Math. Statist., 27 (1956), pp. 642–669.

[64] Rutherford E. and Geiger H.W., Philosophical Magazine, 20 (1910), p. 700.

[65] Schmitt J.H.M.M., Astrophys. J., 293 (1985), p. 178.

[66] Shannon C.E., Proc. IRE, 37 (1949), p. 10.

[67] Student (Gosset W.S.), *On the error of counting with a hæmacitometer*, Biometrika, 5 (1907), pp. 351–360.

[68] ——, *The probable error of the mean*, Biometrika, 6 (1908), pp. 1–25.

[69] Waerden B.L. van der, *Statistique Mathématique*, Dunod, Paris, 1967.

[70] Whittaker E.T., *On the functions which are presented by the expansion of interpolating theory*, Proc. Roy. Soc. Edinburgh A, 35 (1915), pp. 181–194.

[71] Wichmann and Hill, Appl. Stat., 31 Nr. 2 (1982), p. 188.

[72] York D., Can. J. of Phys., 44 (1966), p. 1079.