Statistique Mathématique

F. Balabdaoui-Mohr et O. Wintenberger

Nous tenons à remercier Paul Doukhan et Jean-Marc Bardet pour avoir mis à disposition leurs notes de cours desquelles les chapitres 1, 7 et 8 du présent polycopié sont en grande partie inspirés.

Table des matières

Bibliographie

• Livres pour revoir les bases....

- Baillargeon, B. *Probabilités, statistiques et techniques de régression*. SMG.
- Bercu, B., Pamphile, P. et Azoulay, E. *Probabilités et Applications - Cours Exercices*. Edisciences.
- Dress, F. *Probabilités et Statistique*. Dunod.
- Lecoutre, J.-P. *Statistiques et Probabilités*. Dunod.

• Théorie de la mesure et applications aux probabilités

- Ansel et Ducel, *Exercices corrigés en théorie de la mesure et de l'intégration*, Ellipses.
- Barbe, P. et Ledoux, M., *Probabilités*, Belin.
- Dacunha-Castelle, D. et Duflo, M., *Probabilités et Statistiques (I)*, Masson
- Jacod, J., *Cours d'intégration*, http ://www.proba.jussieu.fr/pageperso/jacod.html.
- Jacod, J., *Cours de Probabilités*, http ://www.proba.jussieu.fr/pageperso/jacod.html.
- Toulouse, P. *Thèmes de probabilités et statistiques*, Masson.

• Statistiques inférentielles

- Dacunha-Castelle, D. et Duflo, M., *Probabilités et Statistiques (I)*, Masson.
- Fourdrinier, D., *Statistique inférentielle*, Dunod.
- Lecoutre, J.-M. et Tassi, P., *Statistique non paramétrique et robustesse*, Economica.
- Milhaud, X., *Statistique*, Belin.
- Monfort, A., *Cours de statistique mathématique*, Economica.
- Saporta, G., *Probabilités, analyse des données et statistiques*. Technip.
- Tsybakov, A. *Introduction à la statistique non-paramétrique*. Collection : Mathématiques et Applications, Springer.

Première partie

Fondements de la statistique mathématique

Chapitre 1

Rappels de probabilités

On commence par des rappels des résultats probabilistes essentiels à la compréhension des méthodes développées dans les chapitres suivants.

1.1 Rappels sur la théorie de la mesure

1.1.1 Mesures

Une mesure est une fonction positive de l'ensemble (tribu) des événements $\mathcal{A}$ d'un espace mesurable $(\Omega, \mathcal{A})$.

• **Tribus**

Dans toute la suite nous adoptons les notations standards :
- Ω est un ensemble (fini ou infini).
- $\mathcal{P}(\Omega)$ est l'ensemble de tous les sous-ensembles (parties) de Ω.

Rappel 1 (Dénombrabilité) *Soit E un ensemble. E est dit dénombrable s'il existe une bijection entre E et $\mathbf{N}$ ou un sous-ensemble de $\mathbf{N}$. Par exemple, un ensemble fini, $\mathbf{Z}$, $\mathbf{D}$, $\mathbf{Z} \times \mathbf{Z}$, $\mathbf{Q}$ sont dénombrables. En revanche, $\mathbf{R}$ n'est pas dénombrable.*

Définition 1.1.1 *Soit une famille $\mathcal{F}$ de parties de Ω (donc $\mathcal{F} \subset \mathcal{P}(\Omega)$). On dit que $\mathcal{F}$ est une algèbre si :*
- *$\Omega \in \mathcal{F}$;*
- *lorsque $A \in \mathcal{F}$ alors le complémentaire $A^c = \overline{A} = (\Omega \setminus A)$ appartient à $\mathcal{F}$;*
- *pour tout $n \in \mathbf{N}^*$, lorsque $(A_1, \cdots, A_n) \in \mathcal{F}^n$ alors la réunion $A_1 \cup \cdots \cup A_n \in \mathcal{F}$.*

Définition 1.1.2 *Soit une famille $\mathcal{A}$ de parties de Ω (donc $\mathcal{A} \subset \mathcal{P}(\Omega)$). On dit que $\mathcal{A}$ est une tribu (ou σ-algèbre) sur Ω si :*
- $\Omega \in \mathcal{A}$;
- *lorsque $A \in \mathcal{A}$ alors $A^c \in \mathcal{A}$;*
- *pour $I \subset \mathbf{N}$, lorsque $(A_i)_{i\in I} \in \mathcal{A}^I$ alors $\bigcup_{i\in I} A_i \in \mathcal{A}$.*

Tout sous ensemble A de Ω élément de la tribu $\mathcal{A}$ est appelé un événement.

Propriété 1 *Avec les notations précédentes :*
1. $\emptyset \in \mathcal{A}$;
2. *si A et B sont dans la tribu $\mathcal{A}$, alors $A \cap B$ est dans $\mathcal{A}$;*
3. *si $\mathcal{A}_1$ et $\mathcal{A}_2$ sont deux tribus sur Ω, alors $\mathcal{A}_1 \cap \mathcal{A}_2$ est une tribu sur Ω . Plus généralement, pour $I \subset \mathbf{N}$, si $(\mathcal{A}_i)_{i\in I}$ ensemble de tribus sur Ω, alors $\bigcap_{i\in I} \mathcal{A}_i$ est une tribu sur Ω ;*
4. *si $\mathcal{A}_1$ et $\mathcal{A}_2$ sont deux tribus sur Ω, alors $\mathcal{A}_1 \cup \mathcal{A}_2$ n'est pas forcément une tribu sur Ω.*

Définition 1.1.3 *Si $\mathcal{E}$ est une famille de parties de Ω (donc $\mathcal{E} \subset \mathcal{P}(\Omega)$), alors on appelle tribu engendrée par $\mathcal{E}$, notée $\sigma(\mathcal{E})$, la tribu engendrée par l'intersection de toutes les tribus contenant $\mathcal{E}$ (on peut faire la même chose avec des algèbres).*

La tribu engendrée est la "plus petite" tribu (au sens de l'inclusion) contenant la famille $\mathcal{E}$.

Rappel 2 (Topologie)
- *Un ensemble ouvert U dans un espace métrique X (muni d'une distance d) est tel que pour tout $x \in U$, il existe $r > 0$ tel que $B(x,r) \subset U$ ou $B(x,r) = \{y \in X;\ d(x,y) < r\}$.*
- *On dit qu'un ensemble dans un espace métrique X est fermé si son complémentaire dans X est ouvert.*

Une tribu naturelle est celle engendrée par les ensembles ouverts (et donc fermés) :

Définition 1.1.4 *Soit Ω un espace métrique. On appelle tribu borélienne sur Ω, notée, $\mathcal{B}(\Omega)$, la tribu engendrée par les ouverts de Ω. Un ensemble de $\mathcal{B}(\Omega)$ est appelé borélien.*

• Espace mesurable

Lorsque Ω est un ensemble et $\mathcal{A}$ une tribu sur Ω on dit que $(\Omega, \mathcal{A})$ est un espace mesurable. Quand on s'intéressera aux probabilités, on dira de même que $(\Omega, \mathcal{A})$ est un espace probabilisable.

Propriété 2 *Si $(\Omega_i, \mathcal{A}_i)_i$ sont n espaces mesurables, alors un ensemble élémentaire de $\Omega = \Omega_1 \times \cdots \times \Omega_n$ est une réunion finie d'ensembles $A_1 \times \cdots \times A_n$ appelés pavés ou cylindres, où chaque $A_i \in \mathcal{A}_i$. L'ensemble des ensembles élémentaires est une algèbre et on note $\mathcal{A}_1 \otimes \cdots \otimes \mathcal{A}_n = \bigotimes_{i=1}^n \mathcal{A}_i$ (on dit $\mathcal{A}_1$ tensoriel $\mathcal{A}_2$ ... tensoriel $\mathcal{A}_n$) la tribu sur Ω engendrée par ces ensembles élémentaires.*

Définition 1.1.5 *On appelle espace mesurable produit des $(\Omega_i, \mathcal{A}_i)_i$ l'espace mesurable $\left(\prod_{i=1}^n \Omega_i, \bigotimes_{i=1}^n \mathcal{A}_i\right)$.*

Il et désormais temps de définir ce qu'est une mesure :

Définition 1.1.6 *Soit $(\Omega, \mathcal{A})$ un espace mesurable. L'application $\mu : \mathcal{A} \to [0, +\infty]$ est une mesure si :*

- *$\mu(\emptyset) = 0$.*
- *Pour tout $I \subset \mathbf{N}$ et pour $(A_i)_{i\in I}$ famille disjointe de $\mathcal{A}$ (telle que $A_i \cup A_j = \emptyset$ pour $i \neq j$), alors $\mu\left(\bigcup_{i\in I} A_i\right) = \sum_{i\in I} \mu(A_i)$ (propriété dite de σ-additivité).*

Une mesure pondère les événements. Avec les notations précédentes :

- Si $\mu(\Omega) < +\infty$, on dit que μ est finie.
- Si $\mu(\Omega) < M$ avec $M < +\infty$, on dit que μ est bornée.
- Si $\mu(\Omega) = 1$, on dit que μ est une mesure de probabilité.

Définition 1.1.7 *Si $(\Omega, \mathcal{A})$ est un espace mesurable (resp. probabilisable) alors $(\Omega, \mathcal{A}, \mu)$ est un espace mesuré (resp. probabilisé quand μ est une probabilité).*

Sur $(\Omega, \mathcal{A})$, on peut définir une infinité de mesures. Elles ont toutes les propriétés suivantes :

Propriété 3 *Soit $(\Omega, \mathcal{A}, \mu)$ un espace mesuré et $(A_i)_{i\in\mathbf{N}}$, une famille de $\mathcal{A}$.*

1. *Si $A_1 \subset A_2$, alors $\mu(A_1) \leq \mu(A_2)$.*
2. *Si $\mu(A_1) < +\infty$ et $\mu(A_2) < +\infty$, alors $\mu(A_1 \cup A_2) + \mu(A_1 \cap A_2) = \mu(A_1) + \mu(A_2)$.*
3. *Pour tout $I \subset \mathbf{N}$, on a $\mu\left(\bigcup_{i\in I} A_i\right) \leq \sum_{i\in I} \mu(A_i)$.*
4. *Si $A_i \subset A_{i+1}$ pour tout $i \in \mathbf{N}$ (suite croissante en sens de l'inclusion), alors $(\mu(A_n))_{n\in\mathbf{N}}$ est une suite croissante et $\mu\left(\bigcup_{i\in\mathbf{N}} A_i\right) = \lim_{i\to+\infty} \mu(A_i)$ (même si cette limite est $+\infty$).*

5. *Si $A_{i+1} \subset A_i$ pour tout $i \in \mathbf{N}$ (suite décroissante en sens de l'inclusion) et $\mu(A_0) < +\infty$, alors $(\mu(A_n))_{n\in\mathbf{N}}$ est une suite décroissante convergente telle que* $\mu\left(\bigcap_{i\in\mathbf{N}} A_i\right) = \lim_{i\to+\infty} \mu(A_i)$.

Définition 1.1.8 *Soit $(\Omega, \mathcal{A}, \mu)$ un espace mesuré et $(A_i)_{i\in\mathbf{N}}$ une famille de $\mathcal{A}$.*

1. *On définit la <u>limite supérieure</u> des événements* $\limsup(A_n)_n = \bigcap_{n\in\mathbf{N}} \bigcup_{m\geq n} A_m$.
   Intuitivement, $\limsup(A_n)_n$ *est l'ensemble des $\omega \in \Omega$ tels que ω appartienne à une infinité de A_n.*
2. *On définit la <u>limite inférieure</u> des événements* $\liminf(A_n)_n = \bigcup_{n\in\mathbf{N}} \bigcap_{m\geq n} A_m$.
   Intuitivement, $\liminf(A_n)_n$ *est l'ensemble des $\omega \in \Omega$ tels que ω appartienne à tous les A_n sauf à un nombre fini d'entre eux.*

On vérifie que lim sup et lim inf sont des événements vérifiant $\mu(\limsup(A_n)_n) = \lim \mu(\bigcup_{m\geq n} A_m)$ et $\mu(\liminf(A_n)_n) = \lim \mu(\bigcap_{m\geq n} A_m)$ les limites étant bien définies car portant sur des suites décroissantes et croissantes d'événements. Ci-dessous un résultat utile à la définition de la mesure de Lebesgue sur $\mathbf{R}$, $\mathbf{R}^n$,...

Théorème 1.1.1 (Théorème d'extension de Hahn - Caratheodory) *Si Ω est un ensemble, $\mathcal{F}$ une algèbre sur Ω, et ν une application de $\mathcal{F}$ dans $[0, +\infty]$ additive (telle que $\nu(A \cup B) = \nu(A) + \nu(B)$ pour $A \cup B = \emptyset$), alors si $\mathcal{A}$ est la tribu engendrée par $\mathcal{F}$, il existe une mesure $\widehat{\nu}$ sur la tribu $\mathcal{A}$ qui coïncide avec ν sur $\mathcal{F}$ (c'est-à-dire que pour tout $F \in \mathcal{F}$, $\widehat{\nu(F)} = \nu(F)$). On dit que $\widehat{\nu}$ prolonge ν sur la tribu $\mathcal{A}$.*

Définition 1.1.9 *Soit $(\Omega, \mathcal{A}, \mu)$ un espace mesuré.*

1. *Pour $A \in \mathcal{A}$, on dit que A est <u>μ-négligeable</u> si $\mu(A) = 0$.*
2. *Soit une propriété $\mathcal{P}$ dépendant des éléments ω de Ω. On dit que $\mathcal{P}$ est vraie <u>μ-presque partout</u> (<u>μ-presque sûrement</u> (p.s.) sur un espace probabilisé) si l'ensemble des ω pour laquelle elle n'est pas vérifiée est μ-négligeable.*

Ainsi la propriété " la suite de fonction $f_n(x) = x^n$ converge vers la fonction $f(x) = 0$" est vraie λ-presque partout sur $[0, 1]$.

• **<u>Fonctions mesurables</u>**

Soit $f : E \mapsto F$, où E et F sont 2 espaces métriques.

– Pour $I \subset F$, on appelle ensemble réciproque de I par f, l'ensemble $f^{-1}(I) = \{x \in E, f(x) \in I\}$.

- (f continue) $\Longleftrightarrow$ (pour tout ouvert U de F alors $f^{-1}(U)$ est un ouvert de E).
- On note $f^{-1}(\mathcal{I})$ l'ensemble de sous-ensembles de Ω tel que $f^{-1}(\mathcal{I}) = \{f^{-1}(I), I \in \mathcal{I}\}$.
- Soit $(\Omega', \mathcal{A}')$ un espace mesurable et soit $f : \Omega \mapsto \Omega'$. Alors $f^{-1}(\mathcal{A})$ est une tribu sur Ω appelée tribu engendrée par f.

Définition 1.1.10 *Soit $(\Omega, \mathcal{A})$ et $(\Omega', \mathcal{A}')$ deux espaces mesurables. Une fonction $f : \Omega \mapsto \Omega'$ est dite mesurable pour les tribus $\mathcal{A}$ et $\mathcal{A}'$ si et seulement si $f^{-1}(\mathcal{A}') \subset \mathcal{A}$ (donc si et seulement si $\forall A' \in \mathcal{A}'$, alors $f^{-1}(A') \in \mathcal{A}$).*

Par exemple, les fonctions indicatrices d'événements et les combinaisons linéaires de telles fonctions indicatrices sont mesurables. Dans le cas où $(\Omega, \mathcal{A})$ est un espace probabilisable, et si $f : \Omega \mapsto \mathbf{R}$, alors si f est une fonction mesurable pour $\mathcal{A}$ et $\mathcal{B}(\mathbf{R})$, alors f est une variable aléatoire. Dans le cas où $(\Omega, \mathcal{A})$ est un espace mesurable, et si $f : \Omega \mapsto (\Omega', \mathcal{B}(\Omega'))$, où Ω' est un espace métrique et $\mathcal{B}(\Omega')$ l'ensemble des boréliens ou tribu borélienne de Ω', si f est une fonction mesurable sur $\mathcal{A}$ et $\mathcal{B}(\Omega')$, alors f est dite fonction borélienne.

Proposition 1.1.1 *Soit $(\Omega, \mathcal{A})$ et $(\Omega', \mathcal{A}')$ deux espaces mesurables et $f : \Omega \mapsto \Omega'$. Soit $\mathcal{F}$ une famille de sous-ensembles de Ω' telle que $\sigma(\mathcal{F}) = \mathcal{A}'$. Alors*

1. *$f^{-1}(\mathcal{F})$ engendre la tribu $f^{-1}(\mathcal{A})$.*
2. *(f mesurable) $\Longleftrightarrow$ ($f^{-1}(\mathcal{F}) \subset \mathcal{A}$)*

En particulier si $(\Omega, \mathcal{A})$ et $(\Omega', \mathcal{A}')$ sont deux espaces mesurables boréliens, alors toute application continue de $\Omega \mapsto \Omega'$ est mesurable. De plus, pour montrer qu'une fonction $f : \Omega \mapsto \mathbf{R}$ est mesurable, il suffit de montrer que la famille d'ensemble $(\{\omega \in \Omega, f(\omega) \leq a\})_{a \in \mathbf{R}} \in \mathcal{A}$.

Propriété 4

1. *Soit f mesurable de $(\Omega, \mathcal{A})$ dans $(\Omega', \mathcal{A}')$ et g mesurable de $(\Omega', \mathcal{A}')$ dans $(\Omega'', \mathcal{A}'')$. Alors $g \circ f$ est mesurable pour $\mathcal{A}$ et $\mathcal{A}'$.*
2. *Soit f_1 mesurable de $(\Omega, \mathcal{A})$ dans $(\Omega_1, \mathcal{A}_1)$ et f_2 mesurable de $(\Omega, \mathcal{A})$ dans $(\Omega_2, \mathcal{A}_2)$. Alors $h : \Omega \mapsto \Omega_1 \times \Omega_2$ telle que $h(\omega) = (f_1(\omega), f_2(\omega))$ est mesurable pour $\mathcal{A}$ et $\mathcal{A}_1 \otimes \mathcal{A}_2$.*
3. *Soit $(f_n)_{n \in \mathbf{N}}$ une suite de fonctions mesurables de $(\Omega, \mathcal{A})$ dans $(\Omega', \mathcal{B}(\Omega'))$, où Ω' est un espace métrique, telle qu'il existe une fonction f limite simple de (f_n) (donc $\forall \omega \in \Omega$, $\lim_{n \to \infty} f_n(\omega) = f(\omega)$). Alors f est mesurable pour $\mathcal{A}$ et $\mathcal{B}(\Omega')$.*

Définition 1.1.11 *Soit f mesurable de $(\Omega, \mathcal{A}, \mu)$ dans $(\Omega', \mathcal{A}')$ et soit $\mu_f : \mathcal{A}' \mapsto [0, +\infty]$ telle que pour tout $A' \in \mathcal{A}'$, on ait $\mu_f(A') = \mu(f^{-1}(A'))$. Alors μ_f est une mesure sur $(\Omega', \mathcal{A}')$ appelée mesure image de μ par f.*

Si μ est une mesure de probabilité et si X est une variable aléatoire alors μ_X est la mesure (loi) de probabilité de la variable aléatoire X.

• **Cas des fonctions réelles mesurables**

Propriété 5 *1. Soit f et g deux fonctions réelles mesurables (de $(\Omega, \mathcal{A}, \mu)$ dans $(\mathbf{R}, \mathcal{B}(\mathbf{R}))$). Alors $\alpha.f$, $f+g$, $\min(f, g)$ et $\max(f, g)$ sont des fonctions réelles mesurables.*

2. Soit $(f_n)_{n\in\mathbf{N}}$ une suite de fonctions réelles mesurables. Alors $\inf(f_n)$ et $\sup(f_n)$ sont des fonctions réelles mesurables.

Définition 1.1.12 *Soit $f : \Omega \to \mathbf{R}$. Alors f est dite étagée s'il existe une famille d'ensembles disjoints $(A_i)_{1\leq i\leq n}$ de Ω et une famille de réels $(\alpha_i)_{1\leq i\leq n}$ telles que pour tout $\omega \in \Omega$, on ait $f(\omega) = \sum_{i=1}^{n} \alpha_i \mathbf{I}_{A_i}(\omega)$.*

Si les A_i sont tous dans $\mathcal{A}$ tribu sur Ω, alors f est $\mathcal{A}$-mesurable.

Théorème 1.1.2 *Toute fonction réelle mesurable à valeurs dans $[0, +\infty]$ est limite simple d'une suite croissante de fonctions étagées.*

On en déduit que si f une fonction réelle mesurable. Alors f est limite simple de fonctions étagées. Pour des calculs sur les fonctions mesurables, il suffit donc de calculer sur les fonctions étagées puis de passer à la limite. C'est la méthode de Lebesgue dans le cas de l'intégration.

1.1.2 Intégration de Lebesgue

Dans toute la suite, on considère $(\Omega, \mathcal{A}, \mu)$ un espace mesuré. On procède par étape pour définir l'intégrale de Lebesgue. L'intégrale de Lebesgue d'une fonction positive est définie à partir de l'intégrale des fonctions indicatrices et du passage à la limite via les fonctions étagées :

1. Soit $f = \mathbf{I}_A$, où $A \in \mathcal{A}$. Alors :

$$\int f \, d\mu = \int_\omega f(\omega) d\mu(\omega) = \mu(A).$$

2. Soit $f = \mathbf{I}_A$, où $A \in \mathcal{A}$ et soit $B \in \mathcal{A}$. Alors :

$$\int_B f\, d\mu = \int_B f(\omega) d\mu(\omega) = \int \mathbf{I}_B \mu(A)(\omega) f(\omega) d\mu(\omega) = \mu(A \cap B).$$

3. Soit f une fonction étagée positive telle que $f = \sum_{i=1}^{n} \alpha_i \mathbf{I}_{A_i}$, où les $A_i \in \mathcal{A}$ et $\alpha_i > 0$ et soit $B \in \mathcal{A}$. Alors :

$$\int_B f\, d\mu = \int_B f(\omega) d\mu(\omega) = \int \mathbf{I}_B(\omega) f(\omega) d\mu(\omega) = \sum_{i=1}^{n} \alpha_i \mu(A_i \cap B).$$

Définition 1.1.13 *Soit f une fonction $\mathcal{A}$-mesurable positive et soit $B \in \mathcal{A}$. Alors l'intégrale de Lebesgue de f par rapport à μ sur B est :*

$$\int_B f\, d\mu = \int \mathbf{I}_B(\omega) f(\omega) d\mu(\omega) = \sup \left\{ \int_B g\, d\mu, \text{ pour } g \text{ étagée positive avec } g \leq f \right\}.$$

Propriété 6 *Soit f une fonction $\mathcal{A}$-mesurable positive et soit A et $B \in \mathcal{A}$. Alors :*

1. *Pour $c \geq 0$, $\int_B cf\, d\mu = c \int_B f\, d\mu$.*
2. *Si $A \subset B$, alors $\int_A f\, d\mu \leq \int_B f\, d\mu$.*
3. *Si g est $\mathcal{A}$-mesurable telle que $0 \leq f \leq g$ alors $0 \leq \int_B f\, d\mu \leq \int_B g\, d\mu$.*
4. *Si $\mu(B) = 0$ alors $\int_B f\, d\mu = 0$.*

Théorème 1.1.3 (Théorème de convergence monotone (Beppo-Lévi)) *Si $(f_n)_n$ est une suite croissante de fonctions mesurables positives convergeant simplement vers f sur Ω, alors :*

$$\lim_{n\to\infty} \left(\int f_n d\mu \right) = \int f\, d\mu = \int \lim_{n\to\infty} f_n d\mu.$$

En particulier pour les séries de fonctions mesurables positives, on peut toujours appliquer le Théorème de convergence monotone et donc inverser la somme et l'intégrale.

Lemme 1.1.1 (Lemme de Fatou) *Soit $(f_n)_n$ est une suite de fonctions mesurables positives alors :*

$$\int \left(\liminf_{n\to\infty} f_n \right) d\mu \leq \liminf_{n\to\infty} \int f_n d\mu.$$

• **Intégrale de Lebesgue d'une fonction réelle et propriétés**

Définition 1.1.14 *Soit $(\Omega, \mathcal{A}, \mu)$ un espace mesuré, $B \in \mathcal{A}$ et soit f une fonction mesurable à valeurs réelles telle que $f = f^+ - f^-$ avec $f^+ = \max(f, 0)$ et $f^- = \max(-f, 0)$. On dit que f est μ-intégrable sur B si $\int_B |f|\, d\mu < +\infty$. On a alors*

$$\int_B f\, d\mu = \int_B f^+\, d\mu - \int_B f^-\, d\mu.$$

Lorsque f est μ-intégrable sur Ω, soit $\int |f|\, d\mu < +\infty$, on note $f \in \mathcal{L}^1(\Omega, \mathcal{A}, \mu)$ (on dit que f est $\mathcal{L}^1$). On a les propriétés de linéarité et de monotonie de l'intégrale : soient f et $g \in \mathcal{L}^1(\Omega, \mathcal{A}, \mu)$. Alors :

1. $\int (\alpha f + \beta g) d\mu = \alpha \int f d\mu + \beta \int g d\mu$ pour $(\alpha, \beta) \in \mathbf{R}^2$.
2. Si $f \leq g$ alors $\int f\, d\mu \leq \int g\, d\mu$.

Théorème 1.1.4 (Théorème de convergence dominée de Lebesgue) *Soit $(f_n)_n$ une suite de fonctions de $\mathcal{L}^1(\Omega, \mathcal{A}, \mu)$ telles que pour tout $n \in \mathbf{N}$, $|f_n| \leq g$ avec $g \in \mathcal{L}^1(\Omega, \mathcal{A}, \mu)$. Si on suppose que (f_n) converge simplement vers f sur Ω alors :*

$$\lim_{n \to \infty} \int f_n d\mu = \int f\, d\mu.$$

On peut étendre le Théorème de Lebesgue dans le cas où $(f_n)_n$ converge presque partout vers f.

Théorème 1.1.5 (Inégalité de Jensen) *Soit $(\Omega, \mathcal{A}, P)$ un espace probabilisé, soit $\phi : \mathbf{R} \mapsto \mathbf{R}$ une fonction convexe et soit $f : \Omega \mapsto \mathbf{R}$ mesurable telle que $\phi(f)$ soit une fonction intégrable par rapport à P. Alors :*

$$\phi\left(\int f\, dP\right) \leq \int \phi(f)\, dP.$$

• **Mesures induites et densités**

Théorème 1.1.6 (Théorème du Transport) *Soit f une fonction mesurable de $(\Omega, \mathcal{A}, \mu)$ dans $(\Omega', \mathcal{A}')$ telle que μ_f soit la mesure induite par f (donc $\mu_f(A') = \mu(f^{-1}(A'))$ pour $A' \in \mathcal{A}'$) et soit ϕ une fonction mesurable de $(\Omega', \mathcal{A}')$ dans $(\mathbf{R}, \mathcal{B}(\mathbf{R}))$. Alors, si $\phi \circ f \in \mathcal{L}^1(\Omega, \mathcal{A}, \mu)$,*

$$\int_{\Omega'} \phi \, d\mu_f = \int_{\Omega} \phi \circ f \, d\mu.$$

Définition 1.1.15 *Soit μ et ν deux mesures sur $(\Omega, \mathcal{A})$. On dit que μ <u>domine</u> ν (ou ν est dominée par μ) et que ν est <u>absolument continue</u> par rapport à μ lorsque pour tout $A \in \mathcal{A}$, $\mu(A) = 0 \Longrightarrow \nu(A) = 0$.*

Soit $(\Omega, \mathcal{A}, \mu)$ un espace mesuré et f une fonction définie sur $(\Omega, \mathcal{A})$ mesurable et positive. On suppose que pour $A \in \mathcal{A}$, $\nu(A) = \int_A f \, d\mu$. Alors, ν est une mesure sur $(\Omega, \mathcal{A})$, dominée par μ. De plus, pour toute fonction g définie sur $(\Omega, \mathcal{A})$ mesurable et positive,

$$\int g d\nu = \int g.f d\mu.$$

Enfin, g est ν intégrable si et seulement si $g.f$ est μ intégrable.

Définition 1.1.16 *On dit que μ mesure sur $(\Omega, \mathcal{A})$ est <u>σ-finie</u> lorsqu'il existe une famille $(A_i)_{i \in I}$, avec I dénombrable, d'ensembles de $\mathcal{A}$ telle que $\bigcup A_i = \Omega$ et $\mu(A_i) < +\infty$ pour tout $i \in I$.*

Théorème 1.1.7 (Théorème de Radon-Nikodym) *On suppose que μ et ν sont deux mesures σ-finies sur $(\Omega, \mathcal{A})$ telles que μ domine ν. Alors il existe une fonction f définie sur $(\Omega, \mathcal{A})$ mesurable et positive, appelée <u>densité</u> de ν par rapport à μ, telle que pour tout $A \in \mathcal{A}$, $\nu(A) = \int_A f \, d\mu$.*

Théorème 1.1.8 (Théorème de Fubini) *Soit $\Omega = \Omega_1 \times \Omega_2$, $\mathcal{A} = \mathcal{A}_1 \otimes \mathcal{A}_2$ et $\mu = \mu_1 \otimes \mu_2$ (mesures σ finies), où $(\Omega_1, \mathcal{A}_1, \mu_1)$ et $(\Omega_2, \mathcal{A}_2, \mu_2)$ sont des espaces mesurés. Soit une fonction $f : \Omega \mapsto \mathbf{R}$, $\mathcal{A}$-mesurable et μ-intégrable. alors :*

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

• **<u>Espaces $\mathcal{L}^p$ et L^p pour $0 < p \leq \infty$</u>**

Soit $(\Omega, \mathcal{A}, \mu)$ un espace mesuré. On appelle espace $\mathcal{L}^p(\Omega, \mathcal{A}, \mu)$, où $p > 0$, l'ensemble des fonctions $f : \Omega \mapsto \mathbf{R}$, mesurables et telles que $\int |f|^p d\mu < +\infty$. Pour $f \in \mathcal{L}^p(\Omega, \mathcal{A}, \mu)$, où $p > 0$, on note $\|f\|_p = \left(\int |f|^p d\mu \right)^{1/p}$. Pour $p = +\infty$ on définit $\|f\|_\infty = \text{essup}_\Omega |f|$ le supremum essentiel de f tel que $\|f\|_\infty \geq |f(\omega)|$ pour presque tout $\omega \in \Omega$. Alors pour tout f mesurable, $f \in \mathcal{L}^\infty(\Omega, \mathcal{A}, \mu)$ lorsque $\|f\|_\infty < \infty$.

Propriété 7 (Inégalité de Hölder) *Soit $p \geq 1$ et $q \geq 1$ tels que $\frac{1}{p} + \frac{1}{q} = 1$, et soit $f \in \mathcal{L}^p(\Omega, \mathcal{A}, \mu)$ et $g \in \mathcal{L}^q(\Omega, \mathcal{A}, \mu)$. Alors, $f\,g \in \mathcal{L}^1(\Omega, \mathcal{A}, \mu)$ et*

$$\|f\,g\|_1 \leq \|f\|_p.\|g\|_q.$$

Propriété 8 (Inégalité de Minkowski) *Soit $p \geq 1$ et soit f et $g \in \mathcal{L}^p(\Omega, \mathcal{A}, \mu)$. Alors, $f + g \in \mathcal{L}^p(\Omega, \mathcal{A}, \mu)$ et*

$$\|f + g\|_p \leq \|f\|_p + \|g\|_p.$$

Pour $p \geq 1$, $\|.\|_p$ définit ainsi sur une semi-norme sur $\mathcal{L}^p(\Omega, \mathcal{A}, \mu)$. Pour obtenir une norme, on se place dans l'espace $L^p(\Omega, \mathcal{A}, \mu)$ obtenu en "quotientant" $\mathcal{L}^p(\Omega, \mathcal{A}, \mu)$ par la relation d'équivalence $f = g$ μ-presque partout (c'est-à-dire que dans $L^p(\Omega, \mathcal{A}, \mu)$ on dira que $f = g$ lorsque $f = g$ μ-presque partout).

Pour f et $g \in L^2(\Omega, \mathcal{A}, \mu)$, on définit le produit scalaire $< f, g >= \int f.g\, d\mu$. On muni ainsi $L^2(\Omega, \mathcal{A}, \mu)$ d'une structure d'espace de Hilbert. On dira que f est orthogonale à g lorsque $< f, g >= 0$. Si A est un sous-espace vectoriel fermé de $L^2(\Omega, \mathcal{A}, \mu)$ (par exemple un sous-espace de dimension finie), alors pour tout $f \in L^2(\Omega, \mathcal{A}, \mu)$, il existe un unique projeté orthogonal de f sur A, noté $P_A(f)$, qui vérifie $P_A(f) = \underset{g \in A}{\text{Arginf}} \|g - f\|_2$.

1.2 Applications de la théorie de la mesure et de l'intégration en Probabilités

1.2.1 Espérance de variables aléatoires

Soit X une variable aléatoire (v.a.) sur $(\Omega, \mathcal{A}, P)$ un espace probabilisé. Pour tout $0 < p < \infty$, $X \in L^p(\Omega, \mathcal{A}, P)$ lorsque $\int |X|^p dP < \infty$ c'est à dire lorsque $\int |x|^p dP_X(x)$ est fini. Alors si $X \in L^1(\Omega, \mathcal{A}, P)$, on définit l'espérance de X par le nombre $\mathbb{E}X = \int X dP = \int x dP_X(x)$. Plus généralement, si $\phi : \mathbf{R} \mapsto \mathbf{R}$ est borélienne et si $\phi(X) \in L^1(\Omega, \mathcal{A}, P)$, on définit l'espérance de $\phi(X)$ par $\mathbb{E}(\phi(X)) = \int \phi(X)\, dP = \int \phi(x)\, dP_X(x)$.

Si X est une variable aléatoire sur $(\Omega, \mathcal{A}, P)$, si $\phi : \mathbf{R} \mapsto \mathbf{R}$ est borélienne telle que $\phi(X) \in L^1(\Omega, \mathcal{A}, P)$, et si P_X est la mesure de probabilité de X alors :

$$\mathbb{E}(\phi(X)) = \int_{\mathbf{R}} \phi(x)\, dP_X(x).$$

Si P_X est absolument continue par rapport à la mesure de Lebesgue (donc X est une v.a. dite continue) alors elle admet une densité f_X, alors $\mathbb{E}(\phi(X)) = \int_{\mathbf{R}} \phi(x) f_X(x) dx$. Si P_X est absolument continue par rapport à la mesure de comptage sur $\mathbf{N}$ (donc X est une v.a. dite discrète), de densité p_X, alors $\mathbb{E}(\phi(X)) = \sum_{k=0}^{\infty} p_X(x_k)\, \phi(x_k)$.

Propriété 9

1. *Soit X et Y des variables aléatoires telles que X et $Y \in L^1(\Omega, \mathcal{A}, P)$. Alors pour tout $(a, b) \in \mathbf{R}^2$, $aX + bY \in L^1(\Omega, \mathcal{A}, P)$ et*
$$\mathbb{E}(aX + bY) = a\mathbb{E}X + b\mathbb{E}Y.$$
2. *Soit X une variable aléatoire sur $(\Omega, \mathcal{A}, P)$, et soit $A \in \mathcal{A}$. Alors $\mathbb{E}(\mathbf{I}_A(X)) = P(X \in A)$.*
3. *Soit X et Y des variables aléatoires telles que $X \in L^p(\Omega, \mathcal{A}, P)$ et $Y \in L^q(\Omega, \mathcal{A}, P)$ avec $1/p + 1/q = 1$ et $p \geq 1$, $q \geq 1$. Alors $XY \in L^1(\Omega, \mathcal{A}, P)$ et*
$$\mathbb{E}|XY| \leq (\mathbb{E}|X|^p)^{1/p} (\mathbb{E}|Y|^q)^{1/q}.$$
4. *Soit X et Y des variables aléatoires telles que X et $Y \in L^p(\Omega, \mathcal{A}, P)$, avec $p \geq 1$. Alors $X + Y \in L^p(\Omega, \mathcal{A}, P)$ et*
$$(\mathbb{E}|X + Y|^p)^{1/p} \leq (\mathbb{E}|X|^p)^{1/p} + (\mathbb{E}|Y|^p)^{1/p}.$$
5. *Soit X une variable aléatoire telle que $X \in L^p(\Omega, \mathcal{A}, P)$ pour $p > 0$. Alors pour tout $0 < r \leq p$, $X \in L^r(\Omega, \mathcal{A}, P)$ et*
$$(\mathbb{E}|X|^r)^{1/r} \leq (\mathbb{E}|X|^p)^{1/p}.$$
6. *Si X est une variable aléatoire sur $(\Omega, \mathcal{A}, P)$, si $\phi : \mathbf{R} \mapsto \mathbf{R}$ est une fonction borélienne convexe telle que X et $\phi(X) \in L^1(\Omega, \mathcal{A}, P)$, alors*
$$\mathbb{E}(\phi(X)) \geq \phi(\mathbb{E}X).$$

Définition 1.2.1 *Pour X et Y des variables aléatoires telles que X et $Y \in L^2(\Omega, \mathcal{A}, P)$, on définit la covariance de X et Y par*
$$Cov(X, Y) = \mathbb{E}\left[(X - \mathbb{E}X)(Y - \mathbb{E}Y)\right];$$
On appelle variance de X, $Var(X) = Cov(X, X) = \mathbb{E}\left[(X - \mathbb{E}X)^2\right] = \mathbb{E}(X^2) - (\mathbb{E}X)^2$.

Sur $L^2(\Omega, \mathcal{A}, P)$, $\mathrm{Cov}\,(.,.)$ définit un produit scalaire. De plus
$$|\mathrm{Cov}\,(X, Y)|^2 \leq Var(X).Var(Y).$$

Il est souvent utile d'utiliser une espérance pour contrôler une probabilité via le résultat :

Proposition 1.2.1 *Soit g une fonction paire, positive et croissante sur $[0, +\infty[$. Alors pour toute v.a. X et $\epsilon > 0$*

$$P(|X| > \epsilon) \leq \frac{\mathbb{E}g(X)}{g(\epsilon)}.$$

Les inégalités suivantes sont des conséquences immédiates de la Proposition 1.2.1 :

- Inégalité de Markov : Soit X une v.a. et $\epsilon > 0$. Alors pour $r > 0$

$$P(|X| > \epsilon) \leq \frac{\mathbb{E}|X|^r}{\epsilon^r}.$$

- Inégalité de Chebychev : Soit X une v.a. d'espérance finie μ et de variance finie σ^2, et $\epsilon > 0$. Alors

$$P(|X - \mu| > \epsilon) \leq \frac{\sigma^2}{\epsilon^2}.$$

1.2.2 Fonction de répartition et quantiles d'une loi de probabilité

Il y a une correspondance bijective entre la connaissance de P_X et celle de la fonction de répartition $F_X(x) = P_X(]-\infty, x])$. La fonction de répartition permet également de définir les quantiles qui sont essentiels à la construction d'intervalles de confiance et de test.

Soit $\alpha \in [0, 1]$. Des propriétés de la fonction de répartition, on en déduit qu'il existe $x_\alpha \in \mathbf{R}$, tel que :

$$\lim_{x \to x_\alpha^-} F_X(x) \leq \alpha \leq F_X(x_\alpha). \tag{1.1}$$

Soit $I_\alpha = \{x_\alpha \in \mathbf{R}$ tel que x_α vérifie (1.1)$\}$. On appelle quantile (ou fractile, ou percentile en anglais) d'ordre α de la loi P_X, noté q_α, le milieu de l'intervalle I_α. Evidemment, lorsque X admet une distribution absolument continue par rapport à la mesure de Lebesgue, $q_\alpha = F_X^{-1}(\alpha)$, où F_X^{-1} désigne la fonction réciproque de F_X.

Trois cas particuliers sont à connaître :

1/ pour $\alpha = 0.5$, $q_{0.5}$ est appelé la médiane de P_X ;

2/ pour $\alpha = 0.25$ et $\alpha = 0.75$ (respectivement), $q_{0.25}$ et $q_{0.25}$ sont appelés premier et troisième quartile (respectivement) de P_X.

3/ pour $\alpha = 0.1, \ldots, 0.9$, on parlera de décile de P_X.

1.2.3 Indépendance

Soit $(\Omega, \mathcal{A}, P)$ un espace probabilisé.

- Soit $(A_i)_{i\in I}$ une famille dénombrable d'événements de $\mathcal{A}$. On dit que les événements $(A_i)_{i\in I}$ sont indépendants si et seulement si pour tous les sous-ensembles finis $K \subset I$,

$$P\left(\bigcap_{i\in K} A_i\right) = \prod_{i\in K} P(A_i).$$

- Soit $(\mathcal{A}_i)_{i\in I}$ une famille de sous-tribus de $\mathcal{A}$ (donc pour tout $i \in I$, $\mathcal{A}_i \subset \mathcal{A}$). On dit que les tribus $(\mathcal{A}_i)_{i\in I}$ sont indépendantes si et seulement si pour tous les sous-ensembles finis $K \subset I$, et pour tous les événements $A_k \in \mathcal{A}_k$ avec $k \in K$, les A_k sont indépendants.
- Soit $(X_i)_{i\in I}$ des variables aléatoires sur $(\Omega, \mathcal{A})$ à valeurs dans $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. On dit que les v.a. $(X_i)_{i\in I}$ sont indépendantes si et seulement si les tribus engendrées $(X_i^{-1}(\mathcal{B}(\mathbb{R})))_{i\in I}$ sont indépendantes.

Proposition 1.2.2 *Si $(X_1, \cdots, X_n)$ sont des variables aléatoires sur $(\Omega, \mathcal{A}, P)$. Alors les (X_i) sont indépendantes si et seulement si $P_{(X_1,\cdots,X_n)} = \bigotimes_{i=1}^{n} P_{X_i}$.*

Proposition 1.2.3 *Si $(X_i)_{i\in I}$ sont des variables aléatoires indépendantes sur $(\Omega, \mathcal{A}, P)$. Alors les (X_i) sont indépendantes si et seulement si pour tout $J \subset I$, J fini, pour toutes fonctions boréliennes $(g_j)_{j\in J}$ telles que $g_j(X_j)$ soit intégrable, alors*

$$\mathbb{E}\left(\prod_{j\in J} g_j(X_j)\right) = \prod_{j\in J} \mathbb{E}(g_j(X_j)).$$

Lemme 1.2.1 (Lemme de Borel-Cantelli) *Soit $(A_n)_{n\in\mathbb{N}}$ une suite d'événements sur $(\Omega, \mathcal{A}, P)$.*

1. *Si $\sum P(A_n) < +\infty$ alors $P(\limsup A_n) = 0$.*
2. *Si les (A_n) sont indépendants, $\sum P(A_n) = +\infty$ implique que $P(\limsup A_n) = 1$.*

1.2.4 Principales lois de probabilités

On rappelle que la fonction Gamma est telle que $\Gamma(a) = \displaystyle\int_0^\infty x^{a-1} \cdot e^{-x}$ pour $a \geq 0$.

• **Loi uniforme discrète**

C'est la loi de probabilité discrète à valeurs dans $\{x_1, \ldots, x_n\}$ telle que

$$P(X = x_i) = \frac{1}{n}.$$

On alors : $\mathbb{E}X = \frac{1}{n}(x_1 + \ldots + x_n)$ et $Var(X) = \frac{1}{n}(x_1^2 + \ldots + x_n^2) - (\mathbb{E}X)^2$.

• **Loi de Bernoulli**

C'est la loi de probabilité discrète notée $\mathcal{B}(p)$ à valeurs dans $\{0, 1\}$ telle que

$$P(X = 1) = p \quad \text{et} \quad P(X = 0) = 1 - p.$$

On alors : $\mathbb{E}X = p$ et $Var(X) = p(1 - p)$.

• **Loi binomiale**

C'est la loi de probabilité discrète notée $\mathcal{B}(n, p)$ à valeurs dans $\{0, 1, \ldots, n\}$ telle que

$$P(X = k) = C_n^k \cdot p^k \cdot (1 - p)^{n-k} \quad \text{pour } k \in \{0, 1, \ldots, n\}.$$

On alors : $X = X_1 + \cdots + X_n$, où (X_i) est une suite de v.a. iid de loi $\mathcal{B}(p)$, d'où $\mathbb{E}X \cdot p$ et $Var(X) \cdot p(1 - p)$.

• **Loi de Poisson**

C'est la loi de probabilité discrète notée $\mathcal{P}(\theta)$ à valeurs dans $\mathbb{N}$ telle que

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

On alors $\mathbb{E}X = \theta$ et $Var(X) = \theta$.

• **Loi uniforme sur** $[a, b]$

Cette loi est généralement notée $\mathcal{U}([a, b])$, où $-\infty < a < b < \infty$. C'est la loi de probabilité à valeurs dans $[a, b]$ de densité par rapport à la mesure de Lebesgue :

$$f_X(x) = \frac{1}{b - a} \cdot \mathbf{1}_{x \in [a,b]}.$$

On a alors $\mathbb{E}X = \frac{b + a}{2}$ et $Var(X) = \frac{(b - a)^2}{12}$.

• **Loi Gamma**

Cette loi est généralement notée $\gamma(p,\theta)$, où $p>0$ et $\theta>0$. C'est la loi de probabilité à valeurs dans $\mathbb{R}_+$ de densité par rapport à la mesure de Lebesgue :

$$f_X(x)=\frac{\theta^p}{\Gamma(p)}\cdot e^{-\theta\cdot x}\cdot x^{p-1}\cdot \mathbf{1}_{x\in\mathbb{R}_+}.$$

On a alors $\mathbb{E}X=\dfrac{p}{\theta}$ et $Var(X)=\dfrac{p}{\theta^2}$.
Si $X\sim\gamma(p,\theta)$ et $Y\sim\gamma(q,\theta)$ avec X et Y indépendantes et $p>0$ et $q>0$, alors $X+Y\sim\gamma(p+q,\theta)$.
Pour $p=1$, la loi $\gamma(p,\theta)$ est la loi exponentielle $\mathcal{E}(\theta)$.

• **Loi Béta**

Cette loi est généralement notée $\beta(p,q)$, où $p>0$ et $q>0$. C'est la loi de probabilité à valeurs dans $[0,1]$ de densité par rapport à la mesure de Lebesgue :

$$f_X(x)=\frac{x^p(1-x)^{q-1}}{B(p,q)}x^{p-1}\cdot\mathbf{1}_{x\in[0,1]},\quad \text{où}\quad B(p,q)=\frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)}.$$

On a alors $\mathbb{E}X=\dfrac{B(p+1,q)}{B(p,q)}$ et $Var(X)=\dfrac{p\cdot q}{(p+q)^2(p+q+1)}$.
Si $X\sim\gamma(p,\theta)$ et $Y\sim\gamma(q,\theta)$ avec X et Y indépendantes et $p>0$ et $q>0$, alors $\dfrac{X}{X+Y}\sim\beta(p,q)$.

• **Loi normale (ou gaussienne) centrée réduite**

Cette loi est généralement notée $\mathcal{N}(0,1)$. C'est la loi de probabilité à valeurs dans $\mathbb{R}$ de densité par rapport à la mesure de Lebesgue :

$$f_X(x)=\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right).$$

On a :

$$\mathbb{E}(X)=0\quad\text{et}\quad Var(X)=1.$$

• **Loi normale (ou gaussienne) de moyenne m et de variance σ^2 :**

Si Z suit la loi $\mathcal{N}(0,1)$, $X=m+\sigma Z$ suit par définition la loi $\mathcal{N}(m,\sigma^2)$, loi normale d'espérance m et de variance σ^2. La densité de X est donnée par :

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

A partir de la loi gaussienne, on peut en déduire les lois suivantes.

• **Loi du χ^2 à n degrés de libertés**

Soit $X_1, \cdots, X_n$, n variables aléatoires indépendantes de loi $\mathcal{N}(0,1)$, alors

$$S = X_1^2 + \cdots + X_n^2$$

suit une loi du χ^2 à n degrés de libertés, loi notée $\chi^2(n)$. Cette loi est à valeurs dans $\mathbb{R}_+$, d'espérance n et de variance $2n$. C'est aussi la loi Gamma $\gamma(n/2, 1/2)$, c'est-à-dire que $X \sim \chi^2(n)$ admet pour densité par rapport à la mesure de Lebesgue :

$$f_X(x) = \frac{1}{2^{n/2} \cdot \Gamma(n/2)} x^{n/2-1} \exp\left(-\frac{x}{2}\right) \cdot \mathbf{1}_{\{x \geq 0\}},$$

Enfin, si X suit une loi $\chi^2(n)$, par définition on dira que $Y = \sigma^2 \cdot X$ suit une loi $\sigma^2 \cdot \chi^2(n)$.

• **Loi de Student à n degrés de libertés**

La loi de Student à n degrés de liberté, notée $T(n)$, est la loi du quotient

$$T = \frac{N}{\sqrt{S/n}}$$

où N suit une loi $\mathcal{N}(0,1)$ et S suit une loi $\chi^2(n)$, N et S étant deux variables aléatoires indépendantes. Il est également possible de déterminer la densité d'une telle loi par rapport à la mesure de Lebesgue, à savoir,

$$f_X(x) = \frac{1}{\sqrt{n} \cdot B(1/2, n/2)} \left(1 + \frac{t^2}{n}\right)^{-(n+1)/2},$$

où la fonction Beta est telle que $B(a,b) = \dfrac{\Gamma(a) \cdot \Gamma(b)}{\Gamma(a+b)}$ pour $a > 0$ et $b > 0$.

Remarque : Par la loi des grands nombres, plus n est grand, plus S est proche de son espérance qui vaut n. Le dénominateur est donc proche de 1. Il s'ensuit que la loi $T(n)$ est d'autant plus proche d'une loi normale que n est grand.

Un des principaux intérêt de la loi de Student réside dans le fait que si $X_1, \cdots, X_n$ sont n variables aléatoires indépendantes de loi $\mathcal{N}(m, \sigma^2)$, si on considère la moyenne et la variance empiriques :

$$\bar{X}_n = \frac{1}{n}(X_1 + \cdots + X_n) \quad \text{et} \quad \bar{\sigma}_n^2 = \frac{1}{n-1}\left((X_1 - \bar{X}_n)^2 + \cdots + (X_d - \bar{X}_n)^2)\right),$$

alors

$$T = \frac{\sqrt{n}(\bar{X}_n - m)}{\sqrt{\bar{\sigma}_n^2}}$$

suit une loi de Student à $(n-1)$ degrés de liberté. Ce résultat découle directement du théorème suivant :

Théorème 1.2.1 (Théorème de Fisher) *Si $X_1, \ldots, X_n$ sont des v.a. iid $\sim \mathcal{N}(\mu, \sigma^2)$, alors les variables aléatoires $\sqrt{n}(\bar{X}_n - \mu)/\sigma \sim \mathcal{N}(0,1)$ et $nS_n^2/\sigma^2 \sim \chi^2_{n-1}$ et elles sont indépendantes.*

• **Loi de Fisher à n_1 et n_2 degrés de liberté**

Soit S_1 et S_2 deux variables aléatoires indépendantes de loi respectives $\chi^2(n_1)$ et $\chi^2(n_2)$. Alors par définition :

$$F = \frac{S_1/n_1}{S_2/n_2}$$

suit une loi de Fisher à n_1 et n_2 degrés de liberté, notée $F(n_1, n_2)$. Si $F \sim F(n_1, n_2)$, alors $\mathbb{E}(F) = \dfrac{n_2}{n_2 - 2}$ lorsque $n_2 > 2$ et $Var(F) = \dfrac{2n_2^2(n_1 + n_2 - 2)}{n_1(n_2 - 4)(n_2 - 2)^2}$ lorsque $n_2 > 4$. De plus, si T suit une loi de Student $T(n)$, alors T^2 suit une loi de Fisher $F(1, n)$

Remarque : Par les mêmes considérations que précédemment, la loi F est d'autant plus proche de 1 que les degrés de liberté n_1 et n_2 sont grands.

1.2.5 Vecteurs aléatoires

Définition 1.2.2 *On dit que X est un vecteur aléatoire sur $(\Omega, \mathcal{A}, P)$, un espace probabilisé, si X est une fonction mesurable de $(\Omega, \mathcal{A})$ dans $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$.*

Lorsque X un vecteur aléatoire sur $(\Omega, \mathcal{A}, P)$ à valeurs dans $\mathbb{R}^d$, la loi (ou mesure) de probabilité de X, P_X, est définie de façon univoque à partir de la fonction de répartition de X, telle que pour $x = (x_1, \cdots, x_d)$,

$$F_X(x) = P_X(\prod_{i=1}^{d}]-\infty, x_i]) = P(X \in \prod_{i=1}^{d}]-\infty, x_i]).$$

Propriété 10 *Soit X un vecteur aléatoire sur $(\Omega, \mathcal{A}, P)$ à valeurs dans $\mathbb{R}^d$. On suppose que $X = (X_1, \cdots, X_d)$. Alors les X_i sont des variables aléatoires sur $(\Omega, \mathcal{A}, P)$, de fonction de répartition*

$$F_{X_i}(x_i) = \lim_{\substack{x_j \to +\infty \\ j \neq i}} F_X(x_1, \cdots, x_i, \cdots, x_d).$$

Les mesures de probabilités P_{X_i} déterminées de façon univoque à partir des F_{X_i} sont appelées lois marginales de X.

On se place maintenant dans la base canonique orthonormale de $\mathbb{R}^d$. Si Z est un vecteur aléatoire à valeurs sur $\mathbb{R}^d$, on définit $\mathbb{E}(Z)$, le vecteur dont les coordonnées sont les espérances des coordonnées de Z. Ainsi, si dans la base canonique de $\mathbb{R}^d$, $Z = (Z_1, \cdots, Z_d)'$,

$$\mathbb{E}(Z) = \mathbb{E}\begin{pmatrix} Z_1 \\ \vdots \\ Z_d \end{pmatrix} = \begin{pmatrix} \mathbb{E}(Z_1) \\ \vdots \\ \mathbb{E}(Z_d) \end{pmatrix}.$$

De la même manière, on définira l'espérance d'une matrice dont les coordonnées sont des variables aléatoires par la matrice dont les coordonnées sont les espérances de chacune de ces variables aléatoires.

Ceci nous permet de définir la matrice de variance-covariance de Z de la manière suivante :

$$Var(Z) = \mathbb{E}\left[(Z - \mathbb{E}(Z)).(Z - \mathbb{E}(Z))'\right]$$

donc si $Z = (Z_1, \cdots, Z_d)'$,

$$Var\begin{pmatrix} Z_1 \\ \vdots \\ Z_d \end{pmatrix} = \begin{pmatrix} Var(Z_1) & \mathrm{Cov}(Z_1, Z_2) & \cdots & \mathrm{Cov}(Z_1, Z_d) \\ \mathrm{Cov}(Z_1, Z_2) & Var(Z_2) & \cdots & \mathrm{Cov}(Z_2, Z_d) \\ \vdots & \vdots & \cdots & \vdots \\ \mathrm{Cov}(Z_1, Z_d) & \mathrm{Cov}(Z_2, Z_d) & \cdots & Var(Z_d) \end{pmatrix}$$

matrice (d, d) dont les éléments diagonaux sont les variances et les éléments non diagonaux sont les covariances des coordonnées de Z (remarquons que la variance de Z_1 est aussi la covariance de Z_1 et de Z_1).

On vérifie également le résultat suivant : si C est une matrice (p, d) à coordonnées constituées de réels constants et si Z est un vecteur aléatoire à valeurs dans $\mathbb{R}^d$, alors $C \cdot Z$ est un vecteur de taille p de matrice de variance-covariance

$$Var(C \cdot Z) = C \cdot Var(Z) \cdot C'.$$

En particulier, si p vaut 1, alors $C = h'$ où h est un vecteur de taille d, et :

$$Var(h' \cdot Z) = h' \cdot Var(Z) \cdot h.$$

Notez que cette dernière quantité est un scalaire. Soit $Y_1, \cdots, Y_d$ des variables aléatoires indépendantes de même loi $\mathcal{N}(0, \sigma^2)$, indépendantes (ce qui, dans le cas gaussien, est équivalent à $\text{Cov}(Y_i, Y_j) = 0$ pour $i \neq j$). On considère le vecteur $Y = (Y_1, \cdots, Y_d)'$. En raison de l'indépendance, Y est un vecteur gaussien admettant une densité f_Y (par rapport à la mesure de Lebesgue sur $\mathbb{R}^d$) qui est le produit des densités de chacune des coordonnées, soit :

$$\begin{aligned} f_Y(y_1, \cdots, y_d) &= f_{Y_1}(y_1) \times f_{Y_2}(y_2) \times \cdots \times f_{Y_d}(y_d) \\ &= \left(2\pi\sigma^2\right)^{-d/2} \exp\left(-\frac{1}{2\sigma^2}(y_1^2 + \cdots + y_d^2)\right) \\ &= \left(2\pi\sigma^2\right)^{-d/2} \exp\left(-\frac{\|y\|^2}{2\sigma^2}\right), \end{aligned}$$

avec $y = (y_1, \cdots, y_d)$. On voit donc que la densité de Y ne dépend que de la norme $\|Y\|$: elle est constante sur toutes les sphères centrées en zéro. Cela implique qu'elle est invariante par rotation ou symétrie orthogonale d'axe passant par 0 : elle est invariante par toutes les isométries de $\mathbb{R}^d$: on dira que Y suit une loi gaussienne isotrope. Rappelons que les isométries correspondent à des changements de bases orthonormées (BON). En conséquence, on a la première propriété importante :

Propriété 11 *Soit Y un vecteur aléatoire de $\mathbb{R}^d$ de loi normale isotrope de variance σ^2, c'est-à-dire que dans une BON les coordonnées de Y vérifient $\mathbb{E}(Y) = 0$ et $Var(Y) = \sigma^2 \cdot Id$. Alors les coordonnées de Y dans toute BON sont encore des lois $\mathcal{N}(0, \sigma^2)$ indépendantes.*

Voici maintenant l'un des résultats (encore appelé Théorème de Cochran) que nous utilisons le plus et nous en donnons donc une démonstration.

Théorème 1.2.2 (Théorème de Cochran) *Soit E_1 et E_2, deux sous-espaces vectoriels orthogonaux de $E = \mathbb{R}^d$ de dimensions respectives k_1 et k_2 et soit Y un vecteur aléatoire de $\mathbb{R}^d$ de loi normale centrée isotrope de variance σ^2. Alors $P_{E_1}(Y)$ et $P_{E_2}(Y)$ sont deux variables aléatoires gaussienne centrées indépendantes et $\|P_{E_1}(Y)\|^2$ (resp. $\|P_{E_2}(Y)\|^2$) est une loi $\sigma^2 \cdot \chi^2(k_1)$ (resp. $\sigma^2 \cdot \chi^2(k_2)$). Ce théorème se généralise naturellement pour $2 < m \leq d$ sous-espaces vectoriels orthogonaux $(E_i)_{1 \leq i \leq m}$ de $E = \mathbb{R}^d$.*

Démonstration : Soit $(e_1, \cdots, e_{k_1})$ et $(e_{k_1+1}, \cdots, e_{k_1+k_2})$ deux BON de E_1 et E_2 (respectivement). L'ensemble de ces deux bases peut être complété en

$$(e_1, \cdots, e_{k_1}, e_{k_1+1}, \cdots, e_{k_1+k_2}, e_{k_1+k_2+1}, \cdots, e_d)$$

pour former une BON de $\mathbb{R}^d$ (du fait que E_1 et E_2 sont orthogonaux).

Soit $(Y_1, \cdots, Y_d)$, les coordonnées de Y dans cette base ; elles sont indépendantes de loi $\mathcal{N}(0, \sigma^2)$ car le changement de base est orthonormal et nous avons vu que la distribution de Y était conservé par transformation isométrique. Comme

$$P_{E_1}(Y) = Y_1 e_1 + \cdots + Y_{k_1} e_{k_1} \Longrightarrow$$
$$\|P_{E_1}(Y)\|^2 = \sigma^2 \left(\left(\frac{Y_1}{\sigma} \right)^2 + \cdots + \left(\frac{Y_{k_1}}{\sigma} \right)^2 \right),$$
$$P_{E_2}(Y) = Y_{k_1+1} e_{k_1+1} + \cdots + Y_{k_1+k_2} e_{k_1+k_2} \Longrightarrow$$
$$\|P_{E_2}(Y)\|^2 = \sigma^2 \left(\left(\frac{Y_{k_1+1}}{\sigma} \right)^2 + \cdots + \left(\frac{Y_{k_1+k_2}}{\sigma} \right)^2 \right).$$

On voit bien ainsi l'indépendance entre les deux projections et le fait que la loi de $\|P_{E_1}(Y)\|^2$ (resp. $\|P_{E_2}(Y)\|^2$) est une loi $\sigma^2 \cdot \chi^2(k_1)$ (resp. $\sigma^2 \cdot \chi^2(k_2)$). □

On peut définir plus généralement un vecteur gaussien Y à valeurs dans $\mathbb{R}^d$ (non dégénéré), d'espérance $\mu \in \mathbb{R}^d$ et de matrice de variance-covariance Σ quelconques (du moment que Σ soit une matrice symétrique définie positive). Cela équivaut à définir un vecteur aléatoire de densité par rapport à la mesure de Lebesgue sur $\mathbb{R}^d$,

$$f_Y(y) = \frac{(2\pi)^{-n/2}}{\det(\Sigma)^{1/2}} \exp\left(-\frac{1}{2} (y-\mu)'.\Sigma^{-1}.(y-\mu) \right),$$

pour $y \in \mathbb{R}^d$, et avec $\det(\Sigma)$ le déterminant de la matrice Σ. Remarquons une nouvelle fois que l'espérance et la variance définissent complètement la loi de probabilité d'un vecteur gaussien.

A partir des propriétés générales sur les vecteurs aléatoires, on obtient le fait que :

Propriété 12 *Soit Y un vecteur gaussien à valeurs dans $\mathbb{R}^d$ (non dégénéré), d'espérance $\mu \in \mathbb{R}^d$ et de matrice de variance-covariance Σ. Soit C une matrice réelle de taille (p, d) où $p \in \mathbb{N}^*$. Alors $C \cdot Y$ est un vecteur gaussien tel que :*

$$C \cdot Y \sim \mathcal{N}(C \cdot \mu \,,\, C \cdot \Sigma \cdot C')$$

On en déduit les conséquences suivantes :

- si Y est un vecteur gaussien isotrope de $\mathbb{R}^d$ de variance σ^2 et h un vecteur de $\mathbb{R}^d$, alors $h' \cdot Y$ est une combinaison linéaire des coordonnées de Y tel que :

$$h' \cdot Y \text{ suit la loi } \mathcal{N}(0, \sigma^2 \cdot h' \cdot h) = \mathcal{N}(0, \sigma^2 \cdot \|h\|^2)$$

– si Y est un vecteur gaussien d'espérance μ et de matrice de variance Σ et si h un vecteur de $\mathbb{R}^d$, alors $h' \cdot Y$ est une combinaison linéaire des coordonnées de Y et :

$$h' \cdot Y \text{ suit la loi unidimensionnelle } \mathcal{N}(h' \cdot \mu, h' \cdot \Sigma \cdot h)$$

1.2.6 Fonctions caractéristiques

Définition 1.2.3 *Soit X un vecteur aléatoire sur $(\Omega, \mathcal{A}, P)$ à valeurs dans $\mathbb{R}^d$. La fonction caractéristique de X est la fonction $\phi_X : \mathbb{R}^d \mapsto \mathbb{C}$ telle que*

$$\phi_X(t) = \mathbb{E}\left[\exp(i < t, X >)\right] = \int_{\mathbb{R}^d} e^{i<t,x>} dP_X(x),$$

où $< . >$ désigne le produit scalaire euclidien sur $\mathbb{R}^d$ tel que $< t, x >= \sum_{i=1}^{d} t_i x_i$ pour $t = (t_1, \cdots, t_d)$ et $x = (x_1, \cdots, x_d)$.

Corollaire 1.2.1 *$(X_1, \cdots, X_n)$ sont des variables aléatoires indépendantes si et seulement si pour tout $(t_1, \cdots, t_n) \in \mathbb{R}^n$,*

$$\phi_{(X_1,\cdots,X_n)}(t_1, \cdots, t_n) = \prod_{j=1}^{n} \phi_{X_j}(t_j).$$

La fonction caractéristique existe sur $\mathbb{R}$ et $\phi_X(0) = 1$. ϕ_X est aussi la transformée de Fourier de la mesure P_X. Elle caractérise complément la loi de X :

Théorème 1.2.3 *Soit X et Y des vecteurs aléatoires sur $(\Omega, \mathcal{A}, P)$ à valeurs dans $\mathbb{R}^d$, de lois P_X et P_Y. Alors $P_X = P_Y$ si et seulement si $\phi_X = \phi_Y$.*

Théorème 1.2.4 (Théorème d'inversion) *Si X est un vecteur aléatoire sur $(\Omega, \mathcal{A}, P)$ à valeurs dans $\mathbb{R}^d$ et si ϕ_X est une fonction intégrable par rapport à la mesure de Lebesgue λ_d sur $\mathbb{R}^d$, alors X admet une densité f_X par rapport à λ_d telle que pour $x \in \mathbb{R}^d$,*

$$f_X(x) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} e^{-i<t,x>} \phi_X(t) dt.$$

Si X est une variable aléatoire sur $(\Omega, \mathcal{A}, P)$ de fonction génératrice ϕ_X. Alors si $\mathbb{E}(|X|^n) < +\infty$ (ou $X \in L^n(\Omega, \mathcal{A}, P)$), ϕ_X est n fois dérivable et $\phi_X^{(n)}(t) = i^n \mathbb{E}(X^n e^{itX})$. On a alors $i^n \mathbb{E}(X^n) = \phi_X^{(n)}(0)$. On retrouve ainsi $\mathbb{E}(X^n)$ le moment d'ordre n de X noté m_n.

1.2.7 Convergence de suites de variables aléatoires

Soit $(X_n)_{n\in\mathbb{N}}$ une suite de variables aléatoires sur $(\Omega, \mathcal{A}, P)$. On dit que

- (X_n) converge en probabilité vers X, noté $X_n \xrightarrow{P} X$, lorsque pour tout $\varepsilon > 0$,
$$\lim_{n\to\infty} P(|X_n - X| > \varepsilon) = 0.$$
- (X_n) converge dans $L^p(\Omega, \mathcal{A}, P)$ vers X, noté $X_n \xrightarrow{L^p} X$, avec $p > 0$, lorsque
$$\lim_{n\to\infty} \|X_n - X\|_p = 0.$$
- (X_n) converge en loi vers X, noté $X_n \xrightarrow{\mathcal{L}} X$, lorsque,
$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x) \text{ pour tout } x \in \mathbb{R} \text{ tel que } F_X \text{ continue en } x.$$
- (X_n) converge presque sûrement vers X, noté $X_n \xrightarrow{p.s.} X$, lorsque
$$P(E) = 1 \text{ où } E = \{\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\}.$$
ou encore lorsque pour tout $\varepsilon > 0$,
$$\lim_{n\to\infty} P(\sup_{m\geq n} |X_m - X| > \varepsilon) = 0.$$

Remarque : Les définitions de la convergence en loi, en probabilité et de la convergence presque sûre se généralisent facilement pour les vecteurs aléatoires de dimension $d > 1$. Pour la convergence en probabilité par exemple, on remplacera $\lim_{n\to+\infty} P(|X_n - X| > \epsilon) = 0$, $\forall\, \epsilon > 0$ par $\lim_{n\to+\infty} P(\|X_n - X\| > \epsilon) = 0$, $\forall\, \epsilon > 0$ où $\|\cdot\|$ est une norme quelconque sur $\mathbb{R}^d$ puisque toutes les normes sont équivalentes sur $\mathbb{R}^d$.

Propriété 13

1. *p.s. et* $L^p \longrightarrow \mathcal{P} \longrightarrow \mathcal{L}$.
2. *pour* $q \geq p$, $L^q \longrightarrow L^p$ *et* $L^\infty \longrightarrow$ *p.s.*
3. *La convergence en loi n'entraîne pas la convergence en probabilité. Mais* $(X_n \xrightarrow{P} C) \Longleftrightarrow (X_n \xrightarrow{\mathcal{L}} C)$ *pour* C *une constante.*
4. *Si* g *est une fonction continue alors* $(X_n \xrightarrow{p.s.} X) \Longrightarrow (g(X_n) \xrightarrow{p.s.} g(X))$, $(X_n \xrightarrow{P} X) \Longrightarrow (g(X_n) \xrightarrow{P} g(X))$ *et* $X_n \xrightarrow{\mathcal{L}} X \Longrightarrow g(X_n) \xrightarrow{\mathcal{L}} g(X)$.
5. *Si pour tout* $\varepsilon > 0$, $\sum_{n=0}^{\infty} P(|X_n - X| > \varepsilon) < +\infty$ *alors* $X_n \xrightarrow{p.s.} X$ *(application du Lemme de Borel-Cantelli).*

6. *Si il existe* $r > 0$ *tel que* $\mathbb{E}(|X_n|^r) < +\infty$ *et* $\sum_{n=0}^{\infty} \mathbb{E}(|X_n - X|^r) < +\infty$ *alors* $X_n \xrightarrow{p.s.} X$.
7. *Si* $X_n \xrightarrow{p.s.} X$ *et* $|X_n|^r \leq Z$ *telle que* Z *est une v.a. positive telle que* $E(Z) < +\infty$, *alors* $X_n \xrightarrow{L_r} X$ *(application du Théorème de la convergence dominée).*

Théorème 1.2.5 (Loi faible des Grands Nombres) *Soit* $(X_n)_{n\in\mathbb{N}}$ *une suite de v.a. iid Alors si* $\mathbb{E}(|X_i|) < +\infty$,

$$\overline{X}_n = \frac{X_1 + \cdots + X_n}{n} \xrightarrow{P} m = \mathbb{E}X_i.$$

Théorème 1.2.6 (Loi forte des Grands Nombres) *Soit* $(X_n)_{n\in\mathbb{N}}$ *une suite de v.a. iid Alors si* $\mathbb{E}(|X_i|) < +\infty$,

$$\overline{X}_n = \frac{X_1 + \cdots + X_n}{n} \xrightarrow{p.s.} m = \mathbb{E}X_i.$$

Théorème 1.2.7 (Théorème de la limite centrale) *Soit* $(X_n)_{n\in\mathbb{N}}$ *une suite de v.a. iid Alors si* $\sigma^2 = \mathbb{E}X_i^2 < +\infty$, *et* $m = \mathbb{E}X_i$,

$$\sqrt{n}\frac{\overline{X}_n - m}{\sigma} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1).$$

Théorème 1.2.8 (Loi forte des Grands Nombres multidimensionnelle) *Soit* $(X_n)_{n\in\mathbb{N}}$ *une suite de vecteurs aléatoires iid à valeurs dans* $\mathbb{R}^d$. *Alors si* $\mathbb{E}(\|X_i\|) < +\infty$ *(pour* $\|.\|$ *une norme sur* $\mathbb{R}^d$*),*

$$\overline{X}_n = \frac{X_1 + \cdots + X_n}{n} \xrightarrow{p.s.} m = \mathbb{E}X_i.$$

Théorème 1.2.9 (Théorème de la limite centrale multidimensionnel) *Soit* $(X_n)_{n\in\mathbb{N}}$ *une suite de vecteurs aléatoires iid à valeurs dans* $\mathbb{R}^d$. *Alors si* Σ *matrice de covariance de chaque* X_i *existe, et* $m = \mathbb{E}X_i$,

$$\sqrt{n}\Big(\overline{X}_n - m\Big) \xrightarrow{\mathcal{L}} \mathcal{N}_d(0,\Sigma).$$

Théorème 1.2.10 (Théorème de Slutsky) *Soit* (A_n, B_n, X_n, X) *des variables aléatoires réelles et* a *et* b *des réels. Supposons que* $A_n \xrightarrow{\mathcal{L}} a$, $B_n \xrightarrow{\mathcal{L}} b$ *et* $X_n \xrightarrow{\mathcal{L}} X$. *Alors*

$$A_nX_n + B_n \xrightarrow{\mathcal{L}} aX + b.$$

Théorème 1.2.11 (δ-méthode) *Soit* $(X_n)_{n\in\mathbb{N}}$ *une suite de vecteurs aléatoires à valeurs dans* $\mathbb{R}^d$, *indépendants et identiquement distribués, telle que* Σ *matrice de covariance de chaque* X_i *existe, et* $m = \mathbb{E}X_i$. *Soit* $g : \mathbb{R}^d \to \mathbb{R}^p$ *une fonction de classe* $\mathcal{C}^1$ *sur un voisinage autour de* m, *de matrice Jacobienne* $J_g(m)$ *en* m. *Alors,*

$$\sqrt{n}\Big(g(\overline{X}_n) - g(m)\Big) \xrightarrow{\mathcal{L}} \mathcal{N}_d(0, J_g(m)\cdot\Sigma\cdot J_g'(m)).$$

1.2.8 Espérance conditionnelle

Définition 1.2.4 *Soit Y une variable aléatoire sur $(\Omega, \mathcal{A}, P)$. Si $\mathcal{B}$ est une sous-tribu de $\mathcal{A}$ et si $Y \in L^2(\Omega, \mathcal{A}, P)$. Alors on note $\mathbb{E}(Y \mid \mathcal{B})$ la projection orthogonale de Y sur $L^2(\Omega, \mathcal{B}, P)$, appelée espérance conditionnelle de Y sachant $\mathcal{B}$. Ainsi :*

$$\mathbb{E}\left|Y - \mathbb{E}(Y \mid \mathcal{B})\right|^2 = \inf_{Z \in L^2(\Omega, \mathcal{B}, P)} \left\{\mathbb{E}\left|Y - Z\right|^2\right\}.$$

Par extension, si $Y \in L^1(\Omega, \mathcal{A}, P)$, on définit l'espérance conditionnelle par rapport à $\mathcal{B}$, comme l'unique (p.s.) variable aléatoire, $\mathcal{B}$-mesurable vérifiant p.s. :

$$\int_B \mathbb{E}(Y \mid \mathcal{B})\, dP = \int_B Y\, dP, \quad \text{pour tout } B \in \mathcal{B}.$$

Par convention, si X un vecteur aléatoire à valeurs dans $\mathbb{R}^n$ sur $(\Omega, \mathcal{A}, P)$ et si Y une variable aléatoire sur $(\Omega, \mathcal{A}, P)$, on note $\mathbb{E}(Y \mid X) = \mathbb{E}(Y \mid X^{-1}(\mathcal{B}(\mathbb{R})))$.

Propriété 14

1. *Lemme de Doob : Pour $Y \in L^1(\Omega, \mathcal{A}, P)$, et X une v.a. de$(\Omega, \mathcal{A}, P)$, alors p.s. $\mathbb{E}(Y \mid X) = h(X)$, avec h une fonction borélienne.*
2. *Pour Y_1 et Y_2 deux variables aléatoires sur $(\Omega, \mathcal{A}, P)$, et $(a, b, c) \in \mathbb{R}^3$, alors*
$$\mathbb{E}(aY_1 + bY_2 + c \mid \mathcal{B}) = a\mathbb{E}(Y_1 \mid \mathcal{B}) + b\mathbb{E}(Y_2 \mid \mathcal{B}) + c.$$
3. *Si $Y_1 \leq Y_2$, alors $\mathbb{E}(Y_1 \mid \mathcal{B}) \leq \mathbb{E}(Y_2 \mid \mathcal{B})$.*
4. *Le Lemme de Fatou, les théorèmes de Beppo-Levi, Lebesgue et Jensen s'appliquent avec l'espérance conditionnelle.*
5. *Si $Y \in L^2(\Omega, \mathcal{B}, P)$, alors $\mathbb{E}(Y \mid \mathcal{B}) = Y$; ainsi $\mathbb{E}(g(X) \mid X) = g(X)$ pour g une fonction mesurable réelle.*
6. *On a $\mathbb{E}(\mathbb{E}(Y \mid \mathcal{B})) = \mathbb{E}Y$.*
7. *Si $Y^{-1}(\mathcal{B}(\mathbb{R}))$ et $\mathcal{B}$ sont indépendantes alors $\mathbb{E}(Y \mid \mathcal{B}) = \mathbb{E}Y$; ainsi, si X et Y sont indépendantes, $\mathbb{E}(Y \mid X) = \mathbb{E}Y$.*
8. *Si (X, Y) est un couple de v.a. à valeurs dans $\mathbb{R}^2$ possédant une densité $f_{(X,Y)}$ par rapport à la mesure de Lebesgue, alors si X est intégrable ,*
$$\mathbb{E}(Y \mid X = x) = \frac{\int_{\mathbb{R}} y \cdot f_{(X,Y)}(x, y)\, dy}{\int_{\mathbb{R}} f_{(X,Y)}(x, y)\, dy},$$
pour tout x tel que $\int_{\mathbb{R}} f_{(X,Y)}(x, y)\, dy > 0$.

Dans le cas Gaussien, on a la propriété suivante :

Proposition 1.2.4 *Si $(Y, X_1, \cdots, X_n)$ est un vecteur gaussien, alors on a*

$$\mathbb{E}(Y \mid (X_1, \cdots, X_n)) = a_0 + a_1 X_1 + \cdots + a_n X_n,$$

où les a_i sont des réels.

Chapitre 2

Échantillonnage

2.1 L'échantillon aléatoire

A partir de l'observation d'une propriété sur un ensemble fini d'expériences, donc d'un ensemble de données de taille finie, le statisticien infère des caractèristiques de la propriété en général. Deux cas de figure sont possibles :

- Soit la propriété est observée sur un sous ensemble de taille n d'une population mère est de taille finie N avec $N \geq n$,
- Soit la propriété est observée sur un ensemble fini d'expériences, ces expériences sont renouvelables théoriquement autant de fois que l'on veut.

On consacre cette section à la notion d'échantillon aléatoire, notion commune aux deux cas de figures.

2.1.1 Population de taille finie

Soit E un ensemble, que nous appellerons population mère, contenant un nombre fini N d'éléments. Nous supposerons que l'on veut étudier une propriété X de cette population. L'objectif serait donc de déterminer les principales caractéristiques de la loi de X.

S'il est possible d'effectuer un recensement, c'est-à-dire interroger ou inspecter tous les éléments de E, les caractéristiques de X seront parfaitement connues. Si on écrit $E = \{e_1, \ldots, e_N\}$ et si X est une propriété mesurable, on observe alors $(x_1, \ldots, x_N)$ l'ensemble des valeurs prises par X correspondant aux éléments de E. Remarquons que ce sont des valeurs déterministes. Dans ce cas précis on peut par exemple calculer les vraies moyenne μ et variance σ^2 de X :

$$\mu = \frac{1}{N}\sum_{j=1}^{N} x_j \qquad \text{et} \qquad \sigma^2 = \frac{1}{N}\sum_{j=1}^{N}(x_j - \mu)^2.$$

Une telle situation est très rare, et l'étude de X sera fréquemment réalisée à partir d'observations partielles de X, ceci pour des considérations de coût, de rapidité de collecte et d'exploitation.

Soit $\mathcal{E}_n$ un échantillon de E de taille n. $\mathcal{E}_n$ est tout simplement un sous-ensemble quelconque de E de n éléments ; $\mathcal{E}_n = \{e_{i_1}, \ldots, e_{i_n}\}$ où $1 \leq i_k \leq N$ et $1 \leq k \leq n$. Il est clair qu'il existe dans ce cas-là C_N^n différentes possibilités pour $\mathcal{E}_n$. Nous supposons ici avoir procédé à la sélection de l'échantillon $\mathcal{E}_n$ de manière aléatoire. On est alors dans le cas d'un tirage aléatoire. Tout calcul statistique sera effectué à partir des valeurs de la propriété X sur l'échantillon choisit aléatoirement $\mathcal{E}_n$. On note $X_1, \ldots, X_n$ les valeurs de X correspondant aux éléments de $\mathcal{E}_n$. Ce sont des variables aléatoires car $\mathcal{E}_n$ a été tiré aléatoirement.

De nombreuses méthodes de tirage aléatoire sont possibles. On étudie ici les deux méthodes suivantes :

- Tirage avec remise : On tire au hasard l'échantillon unité par unité. Lorsqu'un élément est tiré, il n'est pas éliminé. Au contraire, il est "remis" dans la population et peut être tiré ultérieurement. De fait, le même élément peut participer au tirage plusieurs fois. Ce mode de tirage est appelé parfois tirage de Bernoulli
- Tirage sans remise : L'échantillon est obtenu par tirage aléatoire des unités, mains chacune d'entre elles ne peut être tirée qu'une seule fois. Cette méthode d'échantillonnage porte aussi le nom de tirage exhaustif.

2.1.2 Expériences renouvelables

Les modèles de population finie et de tirage aléatoire ne couvrent pas toutes les situations donnant matière à la modélisation statistique. Prenons le cas par exemple de la variable X égale au retard mesuré en minutes que fait le métro d'une ligne quelconque pour arriver à une certaine station. Il est clair que X est une variable aléatoire puisqu'on ne peut exactement prédire le retard (cela dépend de différents facteurs). En revanche, il n'est pas du tout évident comment la notion de population finie et d'échantillonnage aléatoire s'appliquerait ici. On parle plutôt d'expérience que l'on peut renouveler théoriquement autant de fois que l'on veut.

Dans le cas d'expériences renouvelables nous supposerons que celles-ci sont réalisées de la même manière, indépendamment les unes des autres. Dans le cas d'une expérience modélisée par la variable X, alors X_1 correspond à la propriété X mesurée sur la première expérience. L'expérience est renouvelée n fois afin d'obtenir l'échantillon $(X_1, \ldots, X_n)$ puis le statisticien infère à partir de ces données pour déduire des caractéristiques sur X. Ici la vraie loi de X reste inconnu pour toute taille de population. Plus n est grand et plus l'inférence va être bonne.

2.1.3 Modèle d'échantillonnage

Afin de donner à l'échantillonnage un cadre probabiliste général, on utilise la définition suivante.

Définition 2.1.1 *Soit une propriété définie par la v.a. X à valeur dans $\mathcal{X}$, application mesurable de $(\Omega, \mathcal{A}, P) \to (\mathcal{X}, \mathcal{B}, P^X)$, $\mathcal{B}$ étant ici la tribu des Boréliens. Le modèle d'échantillonnage de taille n est l'espace produit*

$$(\mathcal{X}, \mathcal{B}, P)^n = (\mathcal{X}^n, \mathcal{B}_n, P_n^X)$$

où

- $\mathcal{X}^n = \underbrace{\mathcal{X} \times \ldots \times \mathcal{X}}_{n \text{ fois}}$ *est le produit cartésien de l'espace $\mathcal{X}$,*
- $\mathcal{B}_n$ *est la tribu produit des événements de $\mathcal{X}^n$,*
- P_n^X *est la loi ou la distribution jointe des observations.*

On notera X_i la ième observation, v.a. de même loi que X et l'ensemble des observations $(X_1, \ldots, X_n)$ est l'échantillon aléatoire.

Remarque : Dans le cas d'une population de taille finie E, on note $X_1, \ldots, X_n$ les valeurs de X correspondant aux membres de $\mathcal{E}_n$ un échantillon aléatoire de taille $n$de E. La notation $X_1, \ldots, X_n$ ne signifie en aucun cas que les n premiers éléments de la population ont été tirés.

Un cas particulier très important est celui où les observations X_i sont indépendantes entre eux. C'est le cas de la population de taille finie associée à un tirage aléatoire avec remise ou le cas des expériences renouvelables. On notera $X_1, \ldots, X_n$ iid $\sim$ P^X ou iid $\sim$ F_X, F_X étant la fonction de répartition de X. Du fait que les variables aléatoires $X_i, 1 \leq i \leq n$ sont indépendantes, il est facile d'écrire le système de probabilités de l'échantillon $X_1, \ldots, X_n$ dans le cas où la loi P_X est une loi discrète :

$$P(X_1 = x_1, \ldots, X_n = x_n) = \prod_{j=1}^{n} P(X_j = x_j) = \prod_{j=1}^{n} p_X(x_j),$$

ou la densité jointe dans le cas continu (P_X admet une densité f_X relativement à la mesure de Lebesgue) :

$$f_{(X_1,\ldots,X_n)}(x_1, \ldots, x_n) = \prod_{j=1}^{n} f_{X_j}(x_j) = \prod_{j=1}^{n} f_X(x_j).$$

Remarque : Dans le cas de la population finie associée à un tirage sans remise les observations $(X_1, \ldots, X_n)$ sont identiquement distribuées mais pas indépendantes.

2.2 Cas de la population finie

On se place dans le cas d'une population E de taille finie N pour laquelle la propriété X n'est observée que sur un ensemble $\mathcal{E}_n$ de taille $n \leq N$. On note $(x_1, \ldots, x_N)$ l'ensemble des valeurs prises par la propriété X sur l'ensemble de la population $E = \{e_1, \ldots, e_N\}$. Ces valeurs sont déterministes, elles appartiennent à $\mathcal{X}$. On a alors les vraies moyenne μ et variance σ^2 de X :

$$\mu = \frac{1}{N}\sum_{j=1}^{N} x_j \qquad \text{et} \qquad \sigma^2 = \frac{1}{N}\sum_{j=1}^{N}(x_j - \mu)^2.$$

2.2.1 La moyenne empirique

La moyenne empirique de l'échantillon est donnée par l'expression

$$\bar{X}_n = \frac{1}{n}\sum_{j=1}^{n} X_j.$$

Pour calculer $\mathbb{E}(\bar{X}_n)$ et $Var(\bar{X}_n)$ dans le cas d'une population finie E de taille N, il faut distinguer le mode de tirage.

a. Tirage avec remise On a

$$\mathbb{E}(\bar{X}_n) = \frac{1}{n}\sum_{j=1}^{n} \mathbb{E}(X_j).$$

Chacune des variables X_j est tirée de l'ensemble $\{x_1, \ldots, x_N\}$ avec la probabilité $1/N$, c'est-à-dire $P(X_j = x_l) = 1/N, \forall\ l = 1, \ldots, N$. D'où

$$E(X_j) = \frac{1}{N}\sum_{l=1}^{N} x_l = \mu \quad \text{(la vraie moyenne de la population)}$$

et

$$\underline{\mathbb{E}(\bar{X}_n) = \mu.}$$

Pour calculer la variance, notons que les X_j sont des variables aléatoires indépendantes, et donc

$$Var(\bar{X}_n) = \frac{1}{n^2}\sum_{j=1}^{n} Var(X_j)$$

où $\forall\ j = 1, \ldots, n$

$$Var(X_j) = \frac{1}{N}\sum_{l=1}^{N}(x_l - \mu)^2 = \sigma^2 \text{ (la vraie variance de la population).}$$

On en déduit

$$\underline{Var(\bar{X}_n) = \frac{\sigma^2}{n}}.$$

b. Tirage sans remise Dans ce cas, les variables aléatoires X_j ne sont pas indépendantes : Si $X_1 = x_N$ par exemple, on sait qu'il est impossible qu'une des variables $X_j, j \geq 2$ prenne cette valeur aux tirages suivants! Nous devons donc déterminer la loi jointe de l'échantillon :

Proposition 2.2.1 *Dans le cas d'un tirage sans remise, alors $(X_1, \ldots, X_n)$ est équidistribué sur l'ensemble des $n-$arrangements de E. Pour tout $x \in E^n$:*

$$P((X_1, \ldots, X_n) = (x_1, \ldots, x_n)) = \begin{cases} 0 & \text{si il existe } k, k' \text{ tels que } x_k = x_{k'}, \\ \dfrac{(N-n)!}{N!} & \text{sinon.} \end{cases}$$

Démonstration : On raisonne par récurrence sur n. Le résultat est évident pour $n = 1$. Supposons qu'il soit vérifié pour n. Alors on a pour tout $x \in E^{n+1}$:

$$\begin{aligned} P((X_1, \ldots, X_{n+1}) = x) &= P(X_{n+1} = x_{n+1} \,|\, (X_1, \ldots, X_n) = (x_1, \ldots, x_n)) \\ &\quad \cdot P((X_1, \ldots, X_n) = (x_1, \ldots, x_n)). \end{aligned}$$

Intéressons nous aux cas où $(x_1, \ldots, x_n)$ est une combinaison car sinon on trouve 0 d'après l'hypothèse de récurrence. Deux cas de figure sont possible :

- Soit il existe $1 \leq k \leq n$ tels que $i_k = i_{n+1}$ alors $P((X_{n+1} = x_{n+1} \,|\, (X_1, \ldots, X_n) = (x_1, \ldots, x_n)) = 0$ car on est dans le cas d'un tirage sans remise,
- Soit tous les x_i sont différents de x_{n+1}. Alors x est un arrangement étant donné que la valeur x_{n+1} est indissociable des autres $N-n$ de $E\backslash\{x_1, \ldots, x_n\}$. D'où
$$P(X_{n+1} = x_{n+1} \,|\, (X_1, \ldots, X_n) = (x_1, \ldots, x_n)) = \frac{1}{N-n}.$$

On a ainsi facilement le résultat pour $n + 1$. □

On en déduit en particulier que pour tout $1 \leq j \leq n$ et tout $x_l \in E$ avec $l = 1, ..., N$

$$\begin{aligned} P(X_j = x_l) &= \sum_{z \text{ arrangement de } (E\backslash x_l)^{n-1}} P((X_j, (X_i)_{1 \leq i \neq j \leq n}) = (x_l, z)) \\ &= \frac{(N-1)!}{(N-n)!}\frac{(N-n)!}{N!} = \frac{1}{N}. \end{aligned}$$

La loi des X_j restent les mêmes que dans le cas du tirage avec remise et par conséquent

$$\underline{E(\bar{X}_n) = \mu.}$$

Pour la calcul de $Var(\bar{X}_n)$, on utilise la formule générale de la variance de la somme des variables aléatoires dépendantes :

$$Var(\bar{X}_n) = \frac{1}{n^2}\left(\sum_{i=1}^{n} Var(X_i) + \sum_{1\leq j\neq k\leq n} Cov(X_j, X_k)\right).$$

Pour le premier terme, le calcul est donc identique à celui qu'on a déjà effectué pour le tirage avec remise car les X_j ont la même distribution marginale. On trouve que $Var(X_1) = \sigma^2$ et

$$\frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{\sigma^2}{n}$$

Pour le second terme, on a par définition

$$\begin{aligned} Cov(X_i, X_j) &= \mathbb{E}(X_i X_j) - \mathbb{E}(X_i)\mathbb{E}(X_j) \\ &= \mathbb{E}(X_i X_j) - \mu^2. \end{aligned}$$

La détermination des lois jointes des couples (X_i, X_j) pour $i \neq j$ permet de calculer le terme $\mathbb{E}(X_i X_j)$: Pour tout $(x_l, x_{l'}) \in E^2$ avec $l \neq l' \in \{1, ..., N\}^2$

$$\begin{aligned} P((X_i, X_j) = (x_l, x_{l'})) &= \sum_{z \text{ arrangement de } (E\setminus\{x_l, x_{l'}\})^{n-2}} P(((X_i, X_j), (X_k)_{1\leq k\neq i,j\leq n}) = ((x_l, x_{l'}), z) \\ &= \frac{(N-2)!}{(N-n)!}\frac{(N-n)!}{N!} = \frac{1}{N(N-1)}. \end{aligned}$$

On en déduit

$$\begin{aligned} \mathbb{E}(X_i X_j) &= \frac{1}{N(N-1)}\sum_{l\neq l'} x_l x_{l'} = \frac{1}{N(N-1)}\left(\sum_l x_l \sum_{l'} x_{l'} - \sum x_l^2\right) \\ &= \frac{N}{N-1}\mu^2 - \frac{1}{N-1}(\mu^2 + \sigma^2) = \mu^2 - \frac{\sigma^2}{N-1} \end{aligned}$$

et $Cov(X_1, X_2) = -\sigma^2/(N-1)$. Il s'en suit

$$\underline{Var(\bar{X}_n) = \frac{\sigma^2}{n}\frac{N-n}{N-1}.}$$

Remarques :

– En terme de variance, $\bar{X}_n$ est toujours moins dispersé pour un tirage sans remise que pour un tirage avec remise
– Lorsque $N \to +\infty$ et $n/N \to +\infty$, $(N-n)/(N-1) \to 1$ et donc il n'y a pas de différence entre les modes de tirage.

2.2.2 Variance empirique

La variance empirique est donnée par l'expression

$$S_n^2 = \frac{1}{n}\sum_{j=1}^{n}(X_j - \bar{X}_n)^2.$$

S_n^2 peut s'écrire encore

$$S_n^2 = \frac{1}{n}\sum_{j=1}^{n}(X_j - \mu)^2 - (\bar{X}_n - \mu)^2.$$

Par conséquent, dans le cas d'une population finie E de taille N on en déduit

$$\mathbb{E}(S_n^2) = \sigma^2 - Var(\bar{X}_n).$$

a. Tirage avec remise

$$\mathbb{E}(S_n^2) = \frac{n-1}{n}\,\sigma^2.$$

b. Tirage sans remise

$$\mathbb{E}(S_n^2) = \frac{n-1}{n}\,\frac{N}{N-1}\,\sigma^2.$$

La formule de la variance de S_n^2 est plus compliquée à obtenir. Dans le cas d'un tirage avec remise, elle est donnée dans la proposition suivante.

Proposition 2.2.2 *Soit S_n^2 la variance empirique obtenue à partir d'un tirage avec remise. Si $\mathbb{E}[(X-\mu)^4] = \mu_4$ le moment centré d'ordre 4, alors*

$$\begin{aligned} Var(S_n^2) &= \frac{\mu_4 - \sigma^4}{n} - \frac{2(\mu_4 - 2\sigma^2)}{n^2} + \frac{\mu_4 - 3\sigma^4}{n^3} \\ &= \frac{n-1}{n^3}\left((n-1)\mu_4 - (n-3)\sigma^4\right). \end{aligned}$$

Démonstration. Rappelons d'abord que

$$S_n^2 = \frac{1}{n}\sum_{j=1}(X_j - \mu)^2 - (\bar{X}_n - \mu)^2.$$

Posons $Y_j = (X_j - \mu)^2$. On a

$$\begin{aligned} Var(S_n^2) &= \frac{1}{n}Var(Y_1) - \frac{2}{n}\sum_{j=1}^{n} Cov(Y_j, (\bar{X}_n - \mu)^2) + Var((\bar{X}_n - \mu)^2) \\ &= \frac{1}{n}Var(Y_1) - 2Cov(Y_1, (\bar{X}_n - \mu)^2) + Var((\bar{X}_n - \mu)^2) \\ &= I_n - 2II_n + III_n. \end{aligned}$$

$$I_n = \frac{1}{n}(\mathbb{E}[(X_1 - \mu)^4] - \mathbb{E}^2[(X_1 - \mu)^2]) = \frac{\mu_4 - \sigma^4}{n}.$$

D'autre part,

$$\begin{aligned} II_n &= \mathbb{E}[(X_1 - \mu)^2(\bar{X}_n - \mu)^2] - \mathbb{E}[(X_1 - \mu)^2]\mathbb{E}[(\bar{X}_n - \mu)^2] \\ &= \mathbb{E}[(X_1 - \mu)^2(\bar{X}_n - \mu)^2] - \frac{\sigma^4}{n}. \end{aligned}$$

Or,

$$\begin{aligned} \mathbb{E}[(X_1 - \mu)^2(\bar{X}_n - \mu)^2] &= \frac{1}{n^2}\Bigg(\sum_{i=1}^{n}\mathbb{E}[(X_1 - \mu)^2(X_i - \mu)^2] \\ &\qquad + \sum_{j\neq k}\mathbb{E}[(X_1 - \mu)^2(X_j - \mu)(X_k - \mu)]\Bigg) \\ &= \frac{1}{n^2}\Bigg(\mathbb{E}[(X_1 - \mu)^4] + \sum_{i=2}^{n}\mathbb{E}[(X_1 - \mu)^2(X_j - \mu)^2] + 0\Bigg) \\ &= \frac{\mu_4 + (n-1)\sigma^4}{n^2} \end{aligned}$$

d'où

$$\begin{aligned} II_n &= \frac{\mu_4 + (n-1)\sigma^4}{n^2} - \frac{\sigma^4}{n} \\ &= \frac{\mu_4 - \sigma^4}{n^2}. \end{aligned}$$

$$III_n = Var((\bar{X}_n - \mu)^2) = \mathbb{E}[(\bar{X}_n - \mu)^4] - \frac{\sigma^4}{n^2}.$$

Or,

$$\begin{aligned}\mathbb{E}[(\bar{X}_n-\mu)^4] &= \frac{1}{n^4}\Big(\sum_{i=1}^{n}\mathbb{E}[(X_i-\mu)^4]+C_4^2\sum_{j<k}\mathbb{E}[(X_j-\mu)^2(X_k-\mu)^2]+0\Big)\\ &= \frac{n\mu_4+3n(n-1)\sigma^2}{n^4}\\ &= \frac{\mu_4-3\sigma^2}{n^3}+\frac{3\sigma^4}{n^2}.\end{aligned}$$

Il s'en suit que

$$\begin{aligned}Var(S_n^2) &= \frac{\mu_4-\sigma^4}{n}-2\,\frac{\mu_4-\sigma^4}{n^2}+\frac{\mu_4-3\sigma^2}{n^3}+\frac{2\sigma^4}{n^2}\\ &= \frac{\mu_4-\sigma^4}{n}-\frac{2(\mu_4-2\sigma^2)}{n^2}+\frac{\mu_4-3\sigma^4}{n^3}\\ &= \frac{n-1}{n^3}\left((n-1)\mu_4-(n-3)\sigma^4\right). \qquad \square\end{aligned}$$

Remarque : Au premier ordre,

$$Var(S_n^2)\approx\frac{\mu_4-\sigma^4}{n}.$$

Dans le cas d'un tirage sans remise, les calculs deviennent encore plus compliqués. Pour conclure cette section, le cas d'une population de taille finie E comporte de nombreuses difficultés, liés aux calculs de la loi (discrète) de l'échantillon $(X_1,\ldots,X_n)$. Dans toute la suite de ce cours nous nous placerons dans le cadre d'expériences renouvelables qui permet de considérer des lois usuelles ce qui simplifie les calculs.

2.3 Cas d'expériences renouvelables

Soit une expérience qui mesure la propriété X renouvelée n fois. On observe $(X_1,\ldots,X_n)$ iid $\sim P_X$. Le fait que l'expérience puisse être renouveler autant de fois que l'on veut en théorie permet d'utiliser l'asymptotique, à savoir le cas où n tend vers l'infini, et tous les théorèmes de probabilités qui en découle (LFGN, TLC).

2.3.1 Moments empiriques

La LFGN et le TLC nous assure qu'asymptotiquement, lorsque $n\to\infty$, l'approximation de la moyenne μ et la variance σ^2 de la loi X par la moyenne empirique

$\overline{X}_n$ et la variance empirique S_n^2 est bonne. Il est possible de généraliser les notions de moyenne et de variance empiriques, ce qui donne lieu à la notion des moments empiriques.

Définition 2.3.1 *Soit X une variable aléatoire et $(X_1, \ldots, X_n)$ un échantillon de n v.a. indépendantes de même loi que X. Soit $r \in \mathbb{N}^*$.*

- *La quantité*

$$M_n^r = \frac{1}{n}\sum_{j=1}^{n} X_j^r$$

s'appelle le moment empirique d'ordre r.

- *La quantité*

$$\tilde{M}_n^r = \frac{1}{n}\sum_{j=1}^{n} (X_j - \bar{X}_n)^r$$

s'appelle le moment empirique centré d'ordre r.

Moyenne et variance empiriques en sont des cas particuliers :

$$M_n^1 = \bar{X}_n = \frac{1}{n}\sum_{j=1}^{n} X_j, \qquad \tilde{M}_n^2 = S_n^2 = \frac{1}{n}\sum_{j=1}^{n} (X_j - \bar{X}_n)^2.$$

Les moments empiriques ont les propriétés asymptotiques suivantes

Propriété 15

- *Si $\mathbb{E}|X|^r < +\infty$, alors*

$$\begin{aligned} M_n^r &\xrightarrow{p.s.} \mathbb{E}[X^r] = m_r \\ \tilde{M}_n^r &\xrightarrow{p.s.} \mathbb{E}[(X - \mathbb{E}(X))^r] = \mu_r \end{aligned}$$

m_r (resp. μ_r) s'appellent moment (resp. centré) d'ordre r.

- *Si $\mathbb{E}(X^2) < +\infty$, alors*

$$\sqrt{n}(\bar{X}_n - m_1) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \mu_2)$$

- *Si $\mathbb{E}(X^4) < +\infty$, alors*

$$\sqrt{n}(S_n^2 - \mu_2) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \mu_4 - \mu_2^2).$$

2.3.2 Processus empiriques

Nous avons vu comment il était possible d'approcher des caractéristiques de la loi de X telles que les moments etc... Ce sont des quantités appartenant à un espace vectoriel de dimension fini petite. On dit que ces caractéristiques sont paramétriques. On qualifie de non-paramétrique une propriété qui évolue dans un espace fonctionnel de dimension infini, par exemple la fonction de répartition, la fonction quantile, la densité,...

Ici, on se limitera à quelques propriétés asymptotiques élémentaires de la fonction de répartition empirique aussi appelée processus empirique :

Définition 2.3.2 *Soit X une variable aléatoire et $(X_1, \ldots, X_n)$ un $n-$ échantillon de même loi que X. La fonction de répartition empirique F_n est définie par*

$$\mathbb{F}_n(x) = \frac{1}{n} \sum_{j=1}^{n} 1_{X_j \leq x}.$$

Théorème 2.3.1 *Soit $F = F_X$ la fonction de répartition de X. $\forall\ x \in \mathbb{R}$*

$$\begin{aligned} \mathbb{F}_n(x) &\xrightarrow{p.s.} F(x) \\ \sqrt{n}(\mathbb{F}_n(x) - F(x)) &\xrightarrow{\mathcal{L}} \mathcal{N}(0, F(x)(1 - F(x)). \end{aligned}$$

Démonstration : Conséquence immédiate de la loi forte des grands nombres et du théorème central limite. □

2.3.3 Quantiles empiriques

Soit $(X_1, \ldots, X_n)$ un $n-$échantillon d'une v.a. X de fonction de répartition $F_X = F$. Soit α un nombre réel compris entre 0 et 1.

Définition 2.3.3 *Le quantile d'ordre α de X est noté q_α et est donné par la formule*

$$q_\alpha = \inf\{x \in \mathbb{R} \text{ tel que } F_X(x) \geq \alpha\}.$$

Afin d'approcher ce quantile, il faut commencer par ordonner l'échantillon des observations $(X_1, \ldots, X_n)$. Ainsi l'échantillon ordonné dans l'ordre croissant, noté $(X_{(1)}, \ldots, X_{(n)})$ est défini tel que $X_{(k)}$ soit la k-ème plus petite valeur de l'échantillon $(X_1, \ldots, X_n)$.

Définition 2.3.4 *Soit $(X_{(1)}, \ldots, X_{(n)})$ l'échantillon ordonné d'un $n-$échantillon $(X_1, \ldots, X_n)$. On appelle $X_{(j)}$ la statistique d'ordre j.*

Remarques :

– Les statistiques d'ordre ne sont pas indépendantes.
– $X_{(1)} = \min(X_1, \ldots, X_n)$ et $X_{(n)} = \max(X_1, \ldots, X_n)$.

Théorème 2.3.2 (Lois de l'échantillon ordonné) *On suppose X absolument continue et on note $f = f_X$ sa densité. Alors la loi du vecteur $(X_{(1)}, \ldots, X_{(n)})$ a pour densité*

$$\begin{aligned} g_n(z_1, \ldots, z_n) &= n! f(z_1) \ldots f(z_n) \quad si\ z_1 \leq \ldots \leq z_n \\ &= 0 \qquad\qquad\qquad\qquad sinon. \end{aligned}$$

Théorème 2.3.3 (Lois marginales de l'échantillon ordonné) *Si $F = F_X$ est la fonction de répartition de X, alors la fonction de répartition, H_k, de la statistique d'ordre $X_{(k)}$ est donnée par*

$$\begin{aligned} H_k(x) &= \sum_{j=k}^{n} \binom{j}{n} F(x)^j (1 - F(x))^{n-j} \\ &= \frac{n!}{(k-1)!(n-k)!} \int_0^{F(x)} t^{k-1}(1-t)^{n-k} dt. \end{aligned}$$

Si F admet une densité f alors H_k a pour densité

$$h_k(x) = \frac{n!}{(k-1)!(n-k)!} f(x) F(x)^{k-1} (1 - F(x))^{n-k}.$$

Théorème 2.3.4 (Loi jointe de $(X_{(i)}, X_{(j)}), i \neq j$) *On suppose X absolument continue et on note $f = f_X$ sa densité. Alors la loi jointe du couple $(X_{(i)}, X_{(j)}), i < j$ admet pour densité*

$$\begin{aligned} f_{(X_{(i)}, X_{(j)})}(x, y) &= \frac{n!}{(i-1)!(j-i-1)!(n-j)!} F^{i-1}(x) f(x) \times [F(y) - F(x)]^{j-i-1} \\ &\qquad \times (1 - F(y))^{n-j} f(y)\ 1_{x \leq y}. \end{aligned}$$

Le quantile d'ordre α de X va être approcher par une statistique d'ordre appelée quantile empirique :

Définition 2.3.5 *Soit $(X_1, \ldots, X_n)$ un $n-$échantillon de variables aléatoires indépendantes de même loi que X. Le quantile empirique est donné par*

$$\hat{q}_{n,\alpha} = X_{([n\alpha])}$$

où $X_{(1)}, \ldots, X_{(n)}$ sont les statistiques d'ordre croissant et $[y]$ est la partie entière de y.

Proposition 2.3.1 *Soient $\alpha \in]0,1[$ et $\hat{q}_{n,\alpha}$ le quantile empirique d'ordre α. Alors*

$$\hat{q}_{n,\alpha} \xrightarrow{p.s.} q_\alpha$$

où $q_\alpha = F_X^{-1}(\alpha)$ est le quantile théorique d'ordre α.

Démonstration : Par définition de $\hat{q}_{n,\alpha}$ et de la fonction de répartition empirique $\mathbb{F}_n$, on a

$$\mathbb{F}_n(\hat{q}_{n,\alpha}) = \frac{[n\alpha]}{n} \to \alpha \quad \text{lorsque } n \to \infty.$$

Or, par le théorème de Glivenko-Cantelli, on a

$$|\mathbb{F}_n(\hat{q}_{n,\alpha}) - F_X(\hat{q}_{n,\alpha})| \xrightarrow{p.s.} 0$$

ce qui implique

$$F_X(\hat{q}_{n,\alpha}) \xrightarrow{p.s.} \alpha$$

et donc

$$\hat{q}_{n,\alpha} \xrightarrow{p.s.} F_X^{-1}(\alpha) = q_\alpha$$

par continuité de F_X^{-1}. □

Remarque : On peut remplacer $[n\alpha]$ par $[n\alpha]+1$ ou en général par n'importe quelle autre suite k_n pourvu que $k_n/n \to \alpha$. Dans ce cas, l'estimateur sera donné par $X_{(k_n)}$.

Notons maintenant que si $X_1, \ldots, X_n$ sont n v.a. iid de loi F_X alors les v.a. $U_1^*, \ldots, U_n^*$ telles que

$$U_j^* = F_X(X_j)$$

sont n v.a. iid $\sim \mathcal{U}[0,1]$. Cela implique que si $\hat{u}_{n,\alpha}$ est le quantile empirique d'ordre α basé sur un échantillon aléatoire de n v.a. indépendantes de loi uniforme sur $[0,1]$, c'est-à-dire,

$$\hat{u}_{n,\alpha} = U_{(\lfloor \alpha j \rfloor + 1)}$$

où $U_{(1)}, \ldots, U_{(n)}$ sont les statistques d'ordre correspondant à l'échantillon des n variables uniformes $U_1, \ldots, U_n$, alors $\hat{u}_{n,\alpha}$ et $\hat{q}_{n,\alpha}$ ont la même distribution ; i.e.

$$\hat{u}_{n,\alpha} \overset{\mathcal{L}}{=} F_X(\hat{q}_{n,\alpha})$$

Plus généralement, si $k \geq 1$ un entier et $\alpha_1, \ldots, \alpha_k \in]0,1[$ tel que $\alpha_i \neq \alpha_j$ si $i \neq j$, alors

$$\big(\hat{u}_{n,\alpha_1}, \ldots, \hat{u}_{n,\alpha_k}\big) \stackrel{\mathcal{L}}{=} \big(F_X(\hat{q}_{n,\alpha_1}), \ldots, F_X(\hat{q}_{n,\alpha_k})\big).$$

Donc pour établir la loi asymptotique jointe du vecteur aléatoire $(\hat{q}_{n,\alpha_1}, \ldots, \hat{q}_{n,\alpha_k})$, il suffit de le faire pour $(\hat{u}_{n,\alpha_1}, \ldots, \hat{u}_{n,\alpha_k})$. On commence par établir le résultat suivant.

Proposition 2.3.2 *Soient $Y_1, \ldots, Y_{n+1}$ iid $\sim$ Exp(1). Alors les variables aléatoires*

$$Z_j = \frac{Y_1 + \ldots + Y_j}{Y_1 + \ldots + Y_{n+1}}, \quad j = 1, \ldots, n$$

ont la même loi que les statistiques d'ordre d'un échantillon de n v.a. iid de loi uniforme sur $[0,1]$; i.e., si $U_1, \ldots, U_n$ sont iid $\sim \mathcal{U}[0,1]$ alors

$$(Z_1, \ldots, Z_n) \stackrel{\mathcal{L}}{=} (U_{(1)}, \ldots, U_{(n)}).$$

Démonstration : Posons $S_{n+1} = Y_1 + \ldots + Y_{n+1}$ et $Z = (Z_1, \ldots, Z_n)$. La distribution de $(Z_1, \ldots, Z_n, S_{n+1}) = (Z, S_{n+1})$ admet pour densité

$$f_{(Z_1,\ldots,Z_n,S_{n+1})}(z_1, \ldots, z_n, s_{n+1}) = f_{Y_1,\ldots,Y_{n+1}}(y_1, \ldots, y_n, y_{n+1})|J|$$

où J est le Jacobien de la transformation bijective $(y_1, \ldots, y_n, s_{n+1}) \mapsto (z_1, \ldots, z_n, s_{n+1})$:

$$\begin{aligned} y_1 &= s_{n+1} z_1 \\ y_j &= s_{n+1}(z_j - z_{j-1}), \quad j = 2, \ldots, n \\ y_{n+1} &= s_{n+1}(1 - z_n) \end{aligned}$$

et donc $|J| = s_{n+1}^n$. Or, la densité jointe $f_{Y_1,\ldots,Y_n,Y_{n+1}}$ est donnée par

$$f_{Y_1,\ldots,Y_n,Y_{n+1}}(y_1, \ldots, y_n, y_{n+1}) = \prod_{j=1}^{n+1} \exp(-y_j) 1_{y_j \geq 0} = \exp(-s_{n+1}) \prod_{j=1}^{n+1} 1_{y_j \geq 0}.$$

Il s'en suit que

$$\begin{aligned} f_Z(z_1, \ldots, z_n) &= \int_0^\infty \exp(-s_{n+1}) s_{n+1}^n ds_{n+1}\ 1_{z_1 \geq 0} \prod_{j=2}^{n} 1_{z_j \geq z_{j-1}}\ 1_{1 \geq z_n} \\ &= \int_0^\infty \exp(-x) x^n dx\ 1_{z_1 \geq 0} \prod_{j=2}^{n} 1_{z_j \geq z_{j-1}}\ 1_{1 \geq z_n} \\ &= n! \int_0^\infty \frac{1}{n!} \exp(-x) x^{n+1-1} dx\ 1_{0 \leq z_1 \leq \ldots \leq z_n \leq 1} \\ &= n!\ 1_{0 \leq z_1 \leq \ldots \leq z_n \leq 1} \end{aligned}$$

qui est exactement égale à la densité des statistiques d'ordre d'un échantillon de n v.a. iid de loi uniforme sur $[0,1]$. □

Nous énonçons maintenant le théorème principal de cette section.

Théorème 2.3.5 *Soit $(\hat{u}_{n,\alpha_1}, \ldots, \hat{u}_{n,\alpha_k})$ le vecteur des quantiles empiriques d'une loi uniforme sur $[0,1]$. Alors,*

$$\begin{pmatrix} \sqrt{n}(\hat{u}_{n,\alpha_1} - \alpha_1) \\ \sqrt{n}(\hat{u}_{n,\alpha_2} - \alpha_2) \\ \vdots \\ \sqrt{n}(\hat{u}_{n,\alpha_k} - \alpha_k) \end{pmatrix} \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma)$$

où

$$\Sigma = \begin{pmatrix} \alpha_1(1-\alpha_1) & \alpha_1(1-\alpha_2) & \ldots & \alpha_1(1-\alpha_k) \\ \alpha_1(1-\alpha_2) & \alpha_2(1-\alpha_2) & \ldots & \alpha_2(1-\alpha_k) \\ \vdots & \vdots & \vdots & \vdots \\ \alpha_1(1-\alpha_k) & \alpha_2(1-\alpha_k) & \ldots & \alpha_k(1-\alpha_k) \end{pmatrix}.$$

Démonstration : On démontre le théorème pour $k = 2$. Le résultat se généralise sans difficultés au cas général $k > 2$. Posons

$$S_{n,\alpha} = Y_1 + \ldots + Y_{\lfloor n\alpha \rfloor + 1}.$$

La somme $Y_1 + \ldots + Y_{n+1}$ sera encore notée S_{n+1}.

En vue de la proposition 2.3.2, il suffit de montrer que si $Y_1, Y_2, \ldots$ sont des v.a. iid $\sim Exp(1)$, alors

$$\begin{pmatrix} \sqrt{n}\left(\frac{S_{n,\alpha_1}}{S_{n+1}} - \alpha_1\right) \\ \sqrt{n}\left(\frac{S_{n,\alpha_2}}{S_{n+1}} - \alpha_2\right) \end{pmatrix} \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma)$$

où

$$\Sigma = \begin{pmatrix} \alpha_1(1-\alpha_1) & \alpha_1(1-\alpha_2) \\ \alpha_1(1-\alpha_2) & \alpha_2(1-\alpha_2) \end{pmatrix}.$$

Sans perte de généralité, on suppose que $\alpha_2 > \alpha_1$. Par le Théorème Central Limite (on rappelle ici que $E[Y] = 1, Var(Y) = 1$), on a

$$\begin{pmatrix} \sqrt{\lfloor \alpha_1 n \rfloor + 1}\left(\frac{S_{n,\alpha_1}}{\lfloor \alpha_1 n \rfloor} - 1\right) \\ \sqrt{\lfloor \alpha_2 n \rfloor - \lfloor \alpha_1 n \rfloor}\left(\frac{S_{n,\alpha_2} - S_{n,\alpha_1}}{\lfloor \alpha_2 n \rfloor - \lfloor \alpha_1 n \rfloor} - 1\right) \\ \sqrt{n - \lfloor \alpha_2 n \rfloor}\left(\frac{S_{n+1} - S_{n,\alpha_2}}{n - \lfloor \alpha_2 n \rfloor} - 1\right) \end{pmatrix} \xrightarrow{\mathcal{L}} \mathcal{N}(0, I_3)$$

où I_3 est la matrice identité de dimension 3×3. Or, notons que

$$\begin{aligned}
\sqrt{n}\left(\frac{S_{n,\alpha_1}}{n} - \alpha_1\right) &= \sqrt{n}\left(\frac{S_{n,\alpha_1}}{[\alpha_1 n]}\frac{[\alpha_1 n]}{n} - \alpha_1\right) \\
&= \sqrt{n}\left(\frac{S_{n,\alpha_1}}{[\alpha_1 n]} - 1\right) + \sqrt{n}\left(\frac{[\alpha_1 n]}{n} - \alpha_1\right) \\
&= \frac{\sqrt{n}}{\sqrt{[\alpha_1 n]}}\sqrt{[\alpha_1 n]}\left(\frac{S_{n,\alpha_1}}{[\alpha_1 n]} - 1\right) + o(1) \\
&\xrightarrow{\mathcal{L}} \mathcal{N}(0, \alpha_1)
\end{aligned}$$

par le théorème de Slutsky. De la même manière, on peut établir les convergences suivantes

$$\sqrt{n}\left(\frac{S_{n,\alpha_2} - S_{n,\alpha_1}}{n} - (\alpha_2 - \alpha_1)\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \alpha_2 - \alpha_1)$$

et

$$\sqrt{n}\left(\frac{S_{n+1} - S_{n,\alpha_2}}{n} - (1 - \alpha_2)\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1 - \alpha_2).$$

Puisque les variables $S_{n,\alpha_1}, S_{n,\alpha_2} - S_{n,\alpha_1}$ et $S_{n+1} - S_{n,\alpha_2}$ sont indépendantes, il s'en suit que leur loi asymptotique jointe est donnée par

$$\begin{pmatrix} \sqrt{n}\left(\frac{S_{n,\alpha_1}}{n} - \alpha_1\right) \\ \sqrt{n}\left(\frac{S_{n,\alpha_2} - S_{n,\alpha_1}}{n} - (\alpha_2 - \alpha_1)\right) \\ \sqrt{n}\left(\frac{S_{n+1} - S_{n,\alpha_2}}{n} - (1 - \alpha_2)\right) \end{pmatrix} \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma_1)$$

où

$$\Sigma_1 = \begin{pmatrix} \alpha_1 & 0 & 0 \\ 0 & \alpha_2 - \alpha_1 & 0 \\ 0 & 0 & 1 - \alpha_2. \end{pmatrix}$$

Considérons maintenant la transformation

$$g(x, y, z) = \left[\frac{x}{x + y + z}, \frac{x + y}{x + y + z}\right].$$

Alors,

$$g\left(\frac{S_{n,\alpha_1}}{n}, \frac{S_{n,\alpha_2} - S_{n,\alpha_1}}{n}, \frac{S_{n+1} - S_{n,\alpha_2}}{n}\right) = \left[\frac{S_{n,\alpha_1}}{S_{n+1}}, \frac{S_{n,\alpha_2}}{S_{n+1}}\right]$$

et

$$\nabla_{(x,y,z)}g = \frac{1}{(x+y+z)^2}\begin{bmatrix} y+z & -x & -x \\ z & z & -(x+y). \end{bmatrix}.$$

En appliquant la δ-méthode, il s'en suit que

$$\begin{pmatrix} \sqrt{n}\left(\frac{S_{n,\alpha_1}}{S_{n+1}} - \alpha_1\right) \\ \sqrt{n}\left(\frac{S_{n,\alpha_2}}{S_{n+1}} - \alpha_2\right) \end{pmatrix} \xrightarrow{\mathcal{L}} \mathcal{N}(0, G\Sigma_1 G^T)$$

où

$$\begin{aligned} G &= \nabla_{(\alpha_1,\alpha_2-\alpha_1,1-\alpha_2)}g \\ &= \begin{bmatrix} 1-\alpha_1 & -\alpha_1 & -\alpha_1 \\ 1-\alpha_2 & 1-\alpha_2 & -\alpha_2. \end{bmatrix} \end{aligned}$$

Enfin, on peut facilement vérifier que $G\Sigma_1 G^T = \Sigma$. □

Le résultat général, énoncé dans le corollaire suivant, découle immédiatement du théorème précédent et de la δ-méthode.

Corollaire 2.3.1 *Soit f_X la densité de F_X relativement à la mesure de Lebesgue. La loi asymptotique du vecteur $(\hat{q}_{n,\alpha_1}, \ldots, \hat{q}_{n,\alpha_k})$ est donnée par*

$$\begin{pmatrix} \sqrt{n}(\hat{q}_{n,\alpha_1} - q_{\alpha_1}) \\ \sqrt{n}(\hat{q}_{n,\alpha_2} - q_{\alpha_2}) \\ \vdots \\ \sqrt{n}(\hat{q}_{n,\alpha_k} - q_{\alpha_k}) \end{pmatrix} \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma_X)$$

où

$$\Sigma_X = \begin{pmatrix} \frac{\alpha_1(1-\alpha_1)}{f_X^2(q_{\alpha_1})} & \frac{\alpha_1(1-\alpha_2)}{f_X(q_{\alpha_1})f_X(q_{\alpha_2})} & \cdots & \frac{\alpha_1(1-\alpha_k)}{f_X(q_{\alpha_1})f_X(q_{\alpha_k})} \\ \frac{\alpha_1(1-\alpha_2)}{f_X(q_{\alpha_1})f_X(q_{\alpha_2})} & \frac{\alpha_2(1-\alpha_2)}{f_X^2(q_{\alpha_2})} & \cdots & \frac{\alpha_2(1-\alpha_k)}{f_X(q_{\alpha_2})f_X(q_{\alpha_k})} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\alpha_1(1-\alpha_k)}{f_X(q_{\alpha_1})f_X(q_{\alpha_k})} & \frac{\alpha_2(1-\alpha_k)}{f_X(q_{\alpha_2})f_X(q_{\alpha_k})} & \cdots & \frac{\alpha_k(1-\alpha_k)}{f_X^2(q_{\alpha_k})} \end{pmatrix}.$$

Chapitre 3

Exhaustivité et information

Afin de faire de l'inférence statistique, le statisticien va devoir extraire de l'information de l'échantillon $X_1, \ldots, X_n$ dont il dispose. Souvent, ce ne sont pas ces observations elles-mêmes qui seront considérées mais une certaine fonction mesurable de ces données, soit $T_n = T(X_1, \ldots, X_n)$. Ainsi, lorsque la taille de l'échantillon n est grande, il est naturel de tenter de réduire l'échantillon et et de résumer l'information qui y est contenue. Lorsque il est possible de "remplacer" $(X_1, \ldots, X_n)$ par T_n, on optera bien sûr pour cette solution. Dans la suite, on appellera la fonction T une statistique. Par abus, la variable aléatoire $T_n = T(X_1, \ldots, X_n)$ sera aussi appelée statistique.

Cependant, une question se pose : Comment savoir si la réduction des données opérée par la statistique T_n ne conduit pas à une perte d'information? C'est ce type de problèmes que cherche à résoudre la notion d'exhaustivité.

3.1 Exhaustivité

3.1.1 Statistique

Soit X une v.a. à valeurs dans $(\mathcal{X}, \mathcal{B})$ et soit $(\mathcal{Y}, \mathcal{C})$ un espace mesurable auxiliaire quelconque.

Définition 3.1.1 *On appelle statistique toute application T mesurable de $\mathcal{X}^n$ dans $\mathcal{Y}$, $\forall\ n$*

$$T : \mathcal{X}^n \to \mathcal{Y}$$

Par exemple, $\mathcal{X} = \mathcal{Y} = \mathbb{R}$ et

- $T(X_1, \ldots, T_n) = \sum_{j=1}^n X_j/n = \bar{X}_n$.
- $T(X_1, \ldots, T_n) = \sum_{j=1}^n (X_j - \bar{X}_n)^2/n$.

ou $\mathcal{X} = \mathbb{R}$, $\mathcal{Y} = \mathbb{R}^n$ et $T(X_1, \ldots, X_n) = (X_{(1)}, \ldots, X_{(n)})$, où $X_{(1)} \leq X_{(2)} \ldots X_{(n)}$ (cette statistique porte le nom de statistique d'ordre (cf. Chapitre 2).

Remarque : On utilisera la notation $T(X_1, \ldots, X_n)$ pour bien marquer que $T_n = T(X_1, \ldots, X_n)$ est une variable aléatoire.

3.1.2 Statistique exhaustive

Définition 3.1.2 *On appelle modèle statistique paramétrique de paramètre $\theta \in \Theta$ pour un certain espace de dimension fini Θ le couple $(\mathcal{X}, P_\theta)$, où $\mathcal{X}$ est l'espace des valeurs de X, v.a. du modèle, et P_θ la loi de probabilité de X.*

Définition 3.1.3 *La statistique T sera dite exhaustive pour θ si la loi conditionnelle de $\mathbf{X}$ sachant $T(\mathbf{X}) = t$ n'est pas une fonction du paramètre θ :*

$$P_\theta(\mathbf{X} \,|\, T(\mathbf{X}) = t) \quad \textit{ne dépend pas de } \theta.$$

On notera $f(x, \theta)$ la densité de P_θ relativement à une mesure dominante et $\sigma-$finie, ν. On va se restreindre au cas où ν est la mesure de Lebesgue (variables aléatoires de loi absolument continue) et on retrouve la densité $f_\theta(x)$ ou la mesure de comptage (variables aléatoires de loi discrète) et on retrouve le système $P_\theta(X = x)$. On note $\mathbf{X}$ l'échantillon $(X_1, \ldots, X_n)$ issu du même modèle $(\mathcal{X}, P_\theta)$.

Théorème 3.1.1 ***(Théorème de factorisation)*** *Soit le modèle $(\mathcal{X}, P_\theta)$ et T une statistique $(\mathcal{X}^n, \mathcal{B}_n) \to (\mathcal{Y}, \mathcal{C})$. T est exhaustive pour θ si et seulement s'il existe deux fonctions mesurables $g : \mathcal{X} \to \mathbb{R}^+$ et $h : \mathcal{Y} \to \mathbb{R}^+$ telles que $f(\mathbf{x}, \theta)$ se met sous la forme*

$$f(\mathbf{x}, \theta) = h(\mathbf{x}) g(T(\mathbf{x}), \theta)$$

où $\mathbf{x} = (x_1, \ldots, x_n)$.

Exemples :

– Soit $X \sim \mathcal{U}[0, \theta]$. On a

$$f(x_1, \ldots, x_n, \theta) = \frac{1}{\theta^n} 1_{\sup_{1 \leq j \leq n} x_j \leq \theta}.$$

En posant

$$h(\mathbf{x}) = 1 \quad \text{et} \quad g(T(\mathbf{x}), \theta) = \frac{1}{\theta^n} 1_{T(\mathbf{x}) \leq \theta}$$

on déduit que $T : \mathbf{x} \mapsto \sup_{1 \leq j \leq n} x_j$ est une statistique exhaustive pour θ.

– Soit $X \sim Exp(\theta)$. On a

$$f(x_1, \ldots, x_n, \theta) = \frac{1}{\theta^n} \exp\left(-\theta \sum_{j=1}^{n} x_j\right)$$

et donc

$$T(X_1, \ldots, X_n) = \sum_{j=1}^{n} X_j$$

est bien une statistique exhaustive.

– Soit $X \sim \mathcal{P}(\theta)$. On a

$$f(x_1, \ldots, x_n, \lambda) = e^{-n\theta} \frac{\theta^{\sum_{j=1}^{n} x_j}}{\prod_{j=1}^{n} x_j!}$$

et donc

$$T(X_1, \ldots, X_n) = \sum_{j=1}^{n} X_j$$

est bien une statistique exhaustive.

– Soit $X \sim \mathcal{N}(\mu, \sigma^2)$. Alors la statistique

$$T(\mathbf{X}) = \left(\frac{1}{n}\sum_{j=1}^{n} X_j, \frac{1}{n}\sum_{j=1}^{n} X_j^2\right)$$

est une statistique exhaustive pour $\theta = (\mu, \sigma^2)$.

3.1.3 Statistique exhaustive minimale

Le théorème de factorisation implique que si T_1 est une statistique exhaustive pour θ alors la statistique T_2 telle que $T_1 = \varphi \circ T_2$, où φ est une application mesurable, est aussi une statistique exhaustive pour θ. Une statistique exhaustive n'est donc pas unique car il suffit de choisir n'importe quelle application mesurable et bijective, φ, et de considérer $T_2 = \varphi^{-1} \circ T_1$. Ces remarques nous conduisent à la définition suivante.

Définition 3.1.4 *Une statistique T est dite exhaustive <u>minimale</u> pour θ si elle est exhaustive et si pour toute autre statistique exhaustive S pour θ, il existe une application φ telle que $T = \varphi \circ S$.*

Lemme 3.1.1 *Deux statistiques exhaustives minimales pour en θ sont en liaison bijective.*

Soit $f(x,\theta)$ la densité de P_θ par rapport la mesure dominante ν. Le théorème suivant nous donne une condition suffisante pour qu'une statistique soit exhaustive minimale.

Théorème 3.1.2 *Soit une statistique T. Si l'on a l'équivalence*

$$T(x_1,\ldots,x_n) = T(y_1,\ldots,y_n) \;\Leftrightarrow\; \theta \mapsto \frac{f(x_1,\ldots,x_n,\theta)}{f(y_1,\ldots,y_n,\theta)} \text{ ne dépend pas de } \theta$$

alors T est une statistique exhaustive minimale.

Démonstration : Montrons d'abord que T est une statistique exhaustive pour θ. Pour tout $t \in T(\mathcal{X}^n)$, considérons l'ensemble

$$[T=t] = \{\mathbf{x} \in \mathcal{X}^n : T(\mathbf{x}) = t\}\,.$$

à tout élément $\mathbf{x} \in \mathcal{X}^n$, on associe $\mathbf{x}_t$ dans $[T=t]$. On a donc par construction

$$T(\mathbf{x}) = T(\mathbf{x}_{T(\mathbf{x})})$$

et par conséquent, par hypothèse, le rapport

$$h(\mathbf{x}) = \frac{f(\mathbf{x},\theta)}{f(\mathbf{x}_{T(\mathbf{x})},\theta)}$$

est indépendant de θ. Définissons maintenant la fonction $g(T(\mathbf{x}),\theta) = f(\mathbf{x}_{T(\mathbf{x})},\theta)$. On peut écrire

$$f(\mathbf{x},\theta) = h(\mathbf{x})g(T(\mathbf{x}),\theta)$$

et ainsi le théorème de factorisation assure que T est une statistique exhaustive pour θ.

Montrons maintenant que T est minimale. Soit T' une autre statistique exhaustive. Par le théorème de factorisation, il existe deux fonctions h' et g' telles que

$$f(\mathbf{x},\theta) = h'(\mathbf{x})g'(T'(\mathbf{x}),\theta).$$

Alors, pour tout $\mathbf{x}_1$ et $\mathbf{x}_2$ tels que $T'(\mathbf{x}_1) = T'(\mathbf{x}_2)$, il vient que

$$\frac{f(\mathbf{x}_1,\theta)}{f(\mathbf{x}_2,\theta)} = \frac{h'(\mathbf{x}_1)}{h'(\mathbf{x}_2)}.$$

Puisque ce rapport ne dépend pas de θ, l'hypothèse du théorème assure que $T(\mathbf{x}_1) = T(\mathbf{x}_2)$. On en déduit que T doit être une fonction de T' et que T est

donc une statistique exhaustive minimale. □

Exemple : Soient $X_1, \ldots, X_n$ i.i.d. $\mathcal{N}(\mu, \sigma^2)$ où μ et σ sont inconnus. Montrons que $T(X_1, \ldots, X_n) = \bar{X}_n$, la moyenne empirique, est une statistique exhaustive minimale pour μ. Notons tout d'abord que

$$f(x_1, \ldots, x_n, \mu, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{n(\bar{x}_n - \mu)^2 + ns_x^2}{2\sigma^2}\right)$$

où $\bar{x}_n = 1/n \sum_{j=1}^n x_j$ et $s_x^2 = 1/n \sum_{j=1}^n (x_j - \bar{x}_n)^2$. Il s'en suit que

$$\frac{f(x_1, \ldots, x_n, \mu, \sigma^2)}{f(y_1, \ldots, y_n, \mu, \sigma^2)} = \exp\left(-\frac{n(\bar{x}_n - \mu)^2 - n(\bar{y}_n - \mu)^2 + ns_x^2 - ns_y^2}{2\sigma^2}\right).$$

Le rapport ne dépend pas de μ si et seulement si

$$\bar{x}_n = \bar{y}_n.$$

3.2 Statistique libre, complète et notion d'identifiabilité

3.2.1 Statistique libre

Qu'elle serait une sorte d'opposée de la notion de statistique exhaustive minimale ? Ce devrait être une statistique ne dépendant pas du paramètre, soit :

Définition 3.2.1 *Une statistique T d'un modèle paramétrique est dite libre si sa loi ne dépend pas du paramètre θ.*

Une statistique libre n'apporte donc aucune information pour l'estimation du paramètre θ. C'est ce qu'on appelle un paramètre de nuisance. Or, de façon assez surprenante il peut arriver qu'une statistique exhaustive minimale comprenne une statistique libre, qui intuitivement ne devrait pas être prise en compte pour donner toute l'information sur θ. Par exemple la loi P_θ uniforme sur $[\theta, \theta + 1]$; pour un échantillon de taille 2, la statistique $(X_{(2)} - X_{(1)}, X_1 + X_2)$ est exhaustive minimale, mais $X_{(2)} - X_{(1)}$ est libre. Aussi peut-on rajouter une autre caractérisation des statistiques exhaustives pour pouvoir atteindre une forme d'optimalité pour ces statistiques, qui serait qu'aucune fonctionnelle non constante de la statistique ne peut être libre. C'est la notion de statistique complète.

3.2.2 Statistique complète

Définition 3.2.2 *Une statistique exhaustive T d'un modèle statistique paramétrique avec T à valeur dans $\mathbb{R}^d$ est dite complète si pour toute fonction borélienne $f : \mathbb{R}^d \to \mathbb{R}$ telle que $f(T)$ soit intégrable, on ait :*

$$\forall \theta \in \Theta, \quad \mathbb{E}_\theta\big(f(T)\big) = 0 \quad \Longrightarrow \quad f(T) = 0 \qquad P_\theta - p.s.$$

Propriété 16 *Soit un modèle statistique paramétrique dominé.*

1. *si T est une statistique exhaustive complète alors pour toute fonction borélienne φ bijective $\varphi(T)$ est une statistique exhaustive complète.*
2. *si T est une statistique exhaustive complète alors T est une statistique exhaustive minimale.*
3. *(Théorème de Basu) si T est une statistique exhaustive complète alors T est indépendante de toute statistique libre sur le modèle.*

Démonstration : Nous allons prouver le troisième point à savoir le Théorème de Basu. Soit S une statistique libre pour le modèle et soit f une fonction telle que $\mathbb{E}_\theta(f(S))$ existe. On peut noter e l'application linéaire qui à f associe $e(f) = \mathbb{E}_\theta(f(S))$. Comme S est libre, e ne dépend pas de θ. Par suite, et pour tout $\theta' \in \Theta$, la statistique $\mathbb{E}_{\theta'}(f(S) \mid T) - e(f)$ est une fonction de T mesurable telle que $\mathbb{E}_\theta\Big(\mathbb{E}_{\theta'}(f(S) \mid T) - e(f)\Big) = 0$ pour tout $\theta \in \Theta$. Comme on a supposé que T est exhaustive complète, alors $\mathbb{E}_{\theta'}(f(S) \mid T) = e(f)$ presque-sûrement. Autrement dit l'espérance conditionnelle de $f(S)$ par rapport à T est une fonction constante de T. Elle n'est pas aléatoire et les statistiques S et T sont indépendantes. □
Dans un modèle statistique paramétrique, il existe toujours une statistique exhaustive minimale mais pas toujours de statistique exhaustive complète.

3.2.3 Notion d'identifiabilité

Soit $(\mathcal{X}, P_\theta), \theta \in \Theta$ un modèle statistique paramétrique.

Définition 3.2.3 *Une valeur du paramètre $\theta_0 \in \Theta$ est identifiable si $\forall\ \theta \neq \theta_0$, $P_\theta \neq P_{\theta_0}$. Le modèle $(X, P_\theta), \theta \in \Theta$ est dit identifiable si tous les paramètres sont identifiables ; i.e., si l'application $\theta \mapsto P_\theta$ est injective.*

On peut affaiblir la notion précédente à une notion locale.

Définition 3.2.4 *Une valeur du paramètre $\theta_0 \in \Theta$ est localement identifiable s'il existe un voisinage ω_0 de θ_0 tel que $\forall\ \theta \in \omega_0 : \theta \neq \theta_0$ on a $P_\theta \neq P_{\theta_0}$. Le modèle $(X, P_\theta), \theta \in \Theta$ est dit localement identifiable si tous les paramètres sont localement identifiables.*

3.3 Éléments de théorie de l'information

On définira dans cette section différentes quantités mesurant l'information contenue dans un modèle statistique.

3.3.1 Information au sens de Fisher

Soit le modèle statistique $(\mathcal{X}, P_\theta), \theta \in \Theta$ tel que P_θ admet une densité $f(x,\theta)$ relativement à la mesure dominante ν. On appellera hypothèses usuelles les 4 hypothèses suivantes :

H1 : Θ est un ouvert de $\mathbb{R}^d$ pour un certain d fini.

H2 : Le support $\{x : f(x,\theta) > 0\}$ ne dépend pas de θ.

H3 : Pour tout $x \in \mathcal{X}$ la fonction $f(x,\theta)$ est au moins deux fois dérivable par rapport à θ pour tout $\theta \in \Theta$ et que les dérivées première et seconde sont continues. On dit que $\theta \mapsto f(x,\theta)$ est C^2.

H4 : Pour tout $B \in \mathcal{B}$ l'intégrale $\int_B f(x,\theta)d\nu(x)$ est au moins deux fois dérivable sous le signe d'intégration et on peut permuter intégration et dérivation ; i.e.,

$$\begin{aligned} \frac{\partial}{\partial\theta_j}\int_B f(x,\theta)d\nu(x) &= \int_B \frac{\partial f(x,\theta)}{\partial\theta_j}d\nu(x), \quad j = 1,\dots,d \\ \frac{\partial^2}{\partial^2\theta_i\partial\theta_j}\int_B f(x,\theta)d\nu(x) &= \int_B \frac{\partial^2 f(x,\theta)}{\partial\theta_i\partial\theta_j}d\nu(x), \quad i,j \in \{1,\dots,d\}. \end{aligned}$$

Lorsque ces 4 hypothèses sont vérifiées, on dit que le modèle est régulier.

Exemples : Les modèles $X \sim \mathcal{P}(\theta)$, $\theta > 0$, $X \sim Exp(\lambda)$, $\lambda > 0$ et $X \sim \mathcal{N}(\mu,\sigma^2)$, $\theta = (\mu,\sigma^2) \in \mathbb{R}\times\mathbb{R}_+^*$ sont réguliers mais pas $X \sim \mathcal{U}[0,\theta]$, $\theta > 0$.

Définition 3.3.1 *On appelle score le vecteur aléatoire $S(X,\theta)$ définit par*

$$S(X,\theta) = \nabla_\theta(log f(X,\theta)) = \left(\frac{\partial \log f(X,\theta)}{\partial\theta_1},...,\frac{\partial \log f(X,\theta)}{\partial\theta_d}\right)^T.$$

Propriété 17

- *Le score est un vecteur aléatoire centré*

$$\mathbb{E}(S(X,\theta)) = 0.$$

- *Le vecteur score est additif : Soient X et Y deux variables aléatoires indépendantes associées aux modèles statistiques $(\mathcal{X}, P_\theta)$ et $(\mathcal{Y}, Q_\theta)$. Alors $S(X,\theta)$ et $S(Y,\theta)$ sont indépendants*

$$S((X,Y),\theta) = S(X,\theta) + S(Y,\theta),\ \forall\theta \in \Theta.$$

Ici (X,Y) est associé au modèle statistique $(\mathcal{X}\times\mathcal{Y}, P_\theta \otimes Q_\theta)$.

Définition 3.3.2 *On appelle information de Fisher au point θ la matrice*

$$\begin{aligned} I(\theta) &= \mathbb{E}\left[S(X,\theta)S(X,\theta)^T\right] \\ &= \begin{pmatrix} E\left[\left(\frac{\partial \log f(X,\theta)}{\partial \theta_1}\right)^2\right] & E\left[\frac{\partial \log f(X,\theta)}{\partial \theta_1}\frac{\partial \log f(X,\theta)}{\partial \theta_2}\right] & \cdots & E\left[\frac{\partial \log f(X,\theta)}{\partial \theta_1}\frac{\partial \log f(X,\theta)}{\partial \theta_d}\right] \\ \vdots & \vdots & \vdots & \vdots \\ E\left[\frac{\partial \log f(X,\theta)}{\partial \theta_1}\frac{\partial \log f(X,\theta)}{\partial \theta_d}\right] & E\left[\frac{\partial \log f(X,\theta)}{\partial \theta_2}\frac{\partial \log f(X,\theta)}{\partial \theta_d}\right] & \cdots & E\left[\left(\frac{\partial \log f(X,\theta)}{\partial \theta_d}\right)^2\right] \end{pmatrix}. \end{aligned}$$

Théorème 3.3.1 *Pour un modèle régulier, on a la relation*

$$\begin{aligned} I(\theta) &= -E\left[\nabla_\theta (S(X,\theta)^T\right] \\ &= \begin{pmatrix} -E\left[\frac{\partial^2 \log f(X,\theta)}{\partial \theta_1^2}\right] & -E\left[\frac{\partial^2 \log f(X,\theta)}{\partial \theta_1 \partial \theta_2}\right] & \cdots & -E\left[\frac{\partial^2 \log f(X,\theta)}{\partial \theta_1 \partial \theta_d}\right] \\ \vdots & \vdots & \vdots & \vdots \\ -E\left[\frac{\partial^2 \log f(X,\theta)}{\partial \theta_1 \partial \theta_d}\right] & -E\left[\frac{\partial^2 \log f(X,\theta)}{\partial \theta_2 \partial \theta_d}\right] & \cdots & -E\left[\frac{\partial^2 \log f(X,\theta)}{\partial \theta_d^2}\right] \end{pmatrix}. \end{aligned}$$

et donc pour tout $1 \leq i,j \leq d$

$$I_{ij}(\theta) = -E\left[\frac{\partial^2 \log f(X,\theta)}{\partial \theta_i \partial \theta_j}\right].$$

Notons que pour le calcul de $I(\theta)$, l'espérance est prise par rapport à P_θ, à θ fixé.

Propriété 18 *On suppose ici que les hypothèses H1- H4 sont vérifiées, donc que le modèle est régulier.*

- *L'information de Fischer est une matrice symétrique définie positive. En effet, étant donné que le score est centré*

$$I(\theta) = Var(S(X,\theta)) \geq 0.$$

- *L'information de Fischer est additive : Si X et Y deux variables aléatoires indépendantes dans des modèles paramétriques au paramètre θ commun alors*

$$I_{(X,Y)}(\theta) = I_X(\theta) + I_Y(\theta), \quad \forall\ \theta \in \Theta,$$

car c'est la variance d'une somme de scores indépendants.

Exemple. Soit $X \sim \mathcal{N}(\mu, \sigma^2)$, alors

$$I(\mu,\sigma^2) = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix}.$$

En effet,

$$\log f(x,\mu,\sigma^2) = -\frac{1}{2}\log 2\pi - \frac{1}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(x-\mu)^2,$$

$$\frac{\partial^2 \log f(x,\mu,\sigma^2)}{\partial\mu^2} = -\frac{1}{\sigma^2} \Rightarrow -E\left[\frac{\partial^2 \log f(X,\mu,\sigma^2)}{\partial\mu^2}\right] = \frac{1}{\sigma^2}$$

$$\frac{\partial^2 \log f(x,\mu,\sigma^2)}{(\partial\sigma^2)^2} = \frac{1}{2\sigma^4} - \frac{1}{\sigma^6}(x-\mu)^2 \Rightarrow -E\left[\frac{\partial^2 \log f(X,\mu,\sigma^2)}{(\partial\sigma 2)^2}\right] = \frac{1}{2\sigma^4}$$

$$\frac{\partial^2 \log f(x,\mu,\sigma^2)}{\partial\mu\partial\sigma^2} = 0 \Rightarrow E\left[\frac{\partial^2 \log f(X,\mu,\sigma^2)}{\partial\mu\partial\sigma^2}\right] = 0.$$

Pour un échantillon $X_1,\ldots,X_n$, le vecteur score $S_n(\theta)$ et l'information de Fischer $I_n(\theta)$ associés à sont donnés par

$$S_n(\theta) = \nabla_\theta\left(\sum_{i=1}^n \log f(X_i,\theta)\right) \qquad \text{et} \qquad I_n(\theta) = Var(S_n(\theta)).$$

On déduit de l'indépendance des X_j que

$$S_n(\theta) = \sum_{j=1}^n S(X_j,\theta)$$

où les scores $S(X_1,\theta),\ldots,S(X_n,\theta)$ sont i.i.d. (la loi de $S(X,\theta)$ est l'image de la loi de X par l'application $S : x \mapsto S(x,\theta)$). Etant donné que

$$E(S(X,\theta)) = 0, \quad \text{et} \quad Var(S(X,\theta)) = I(\theta) < +\infty,$$

on a donc la relation

$$I_n(\theta) = nI(\theta).$$

En vertu de la loi forte des grands nombres et du théorème central limite, on a aussi :

$$\frac{1}{n}S_n(\theta) \xrightarrow{p.s.} 0 \qquad \text{et} \qquad \frac{S_n(\theta)}{\sqrt{n}} \xrightarrow{\mathcal{L}} \mathcal{N}_d\left(0, I(\theta)\right).$$

3.3.2 Information au sens de Kullback

Soit le modèle statistique $(X, P_\theta), \theta \in \Theta$, et soit θ_0 la vraie valeur (inconnue) du paramètre θ.

Définition 3.3.3 *On appelle pouvoir discriminant de θ_0 contre $\theta_1 \in \Theta$ la quantité*

$$\log\left(\frac{f(x,\theta_0)}{f(x,\theta_1)}\right).$$

Définition 3.3.4 *On appelle information de Kullback la quantité $K(\theta_0, \theta_1)$:*

$$K(\theta_0,\theta_1) = E_{\theta_0}\left[\log\frac{f(X,\theta_0)}{f(X,\theta_1)}\right]$$

La notation E_{θ_0} veut dire que l'espérance est prise par rapport au modèle P_{θ_0} : Si $f(x,\theta_0)$ est la densité de P_{θ_0} par la mesure ν, alors

$$K(\theta_0,\theta_1) = \int \log\left[\frac{f(x,\theta_0)}{f(x,\theta_1)}\right] f(x,\theta_0) d\nu(x).$$

Remarque : $K(\theta_0, \theta_1)$ est appelée aussi la divergence Kullback-Leibler entre les densités $f(x,\theta_0)$ et $f(x,\theta_1)$. Elle vérifie tous les axiomes d'une distance à part la symétrie.

Propriété 19

- $K(\theta_0,\theta_1) \geq 0$. *En effet, d'après l'inégalité de Jensen, étant donné que le logarithme est concave :*

$$\begin{aligned}\mathbb{E}_{\theta_0}\log\left[\frac{f(X,\theta_1)}{f(X,\theta_0)}\right] &\leq \log E_{\theta_0}\left[\frac{f(X,\theta_1)}{f(X,\theta_0)}\right] \\ &= \log\left[\int \frac{f(x,\theta_1)}{f(x,\theta_0)} f(x,\theta_0) d\nu(x)\right] \\ &= \log 1 = 0.\end{aligned}$$

- $K(\theta_0,\theta_1) = 0 \Leftrightarrow f(x,\theta_0) \overset{p.s.}{=} f(x,\theta_1)$.

Exemple : Soit $X \sim \mathcal{P}(\theta)$.

$$\frac{f(x,\theta_0)}{f(x,\theta_1)} = \exp(-(\theta_0-\theta_1))\left[\frac{\theta_0}{\theta_1}\right]^x \Rightarrow \log\frac{f(x,\theta_0)}{f(x,\theta_1)} = \theta_1 - \theta_0 + x\log\frac{\theta_0}{\theta_1}$$

$$K(\theta_0,\theta_1) = \theta_1 - \theta_0 + \theta_0\log\frac{\theta_0}{\theta_1}.$$

Théorème 3.3.2 *Soit $\theta \in \mathbb{R}$. Si les hypothèses H1-H3 de la sous-section 3.3.1 sont vérifiées et si en plus on peut dériver $K(\theta_0, \theta)$ au moins deux fois par rapport à θ sous le signe d'intégration, alors*

$$\frac{\partial^2}{\partial\theta_i \partial\theta_j} K(\theta_0, \theta)|_{\theta=\theta_0} = I_{ij}(\theta_0)$$

où $I(\theta_0)$ est l'information de Fisher au point θ_0.

Démonstration : On a

$$\begin{aligned} \frac{\partial^2}{\partial\theta_i \partial\theta_j} K(\theta_0, \theta)|_{\theta=\theta_0} &= -\int \frac{\partial^2}{\partial\theta_i \partial\theta_j} S(x, \theta)|_{\theta=\theta_0} f(x, \theta_0) d\nu(x) \\ \Rightarrow \frac{\partial^2}{\partial\theta_i \partial\theta_j} K(\theta_0, \theta)|_{\theta=\theta_0} &= -E_{\theta_0}\left[\frac{\partial^2}{\partial\theta_i \partial\theta_j} S(X, \theta)|_{\theta=\theta_0}\right] = I_{ij}(\theta_0). \quad \square \end{aligned}$$

3.3.3 Information et exhaustivité

Le résultat suivant établit le lien étroit qui existe entre les notions de statistique exhaustive et d'information. On se restreint ici au cas unidimensionnel. Soit $I_n(\theta) = nI(\theta)$ l'information de Fisher contenue dans le modèle $(\mathcal{X}^n, P_\theta^n), \theta \in \Theta$. Considérons T une statistique $(\mathcal{X}^n, \mathcal{B}_n, P_\theta^n) \to (\mathcal{Y}, \mathcal{C}, P_\theta^T)$ où P_θ^T est la loi image de P_θ par la transformation T. On note $I_T(\theta)$ l'information contenue dans le modèle image $(\mathcal{Y}, P_\theta^T)$. On rappelle que pour deux matrices A et B on a $A \leq B \Leftrightarrow B - A$ est une matrice symétrique positive.

Théorème 3.3.3 *Pour toute statistique T on a*

$$I_T(\theta) \leq I_n(\theta)$$

et

$$I_T(\theta) = I_n(\theta) \Leftrightarrow \ T \text{ est exhaustive}, \qquad I_T(\theta) = 0 \Leftrightarrow \ T \text{ est libre}.$$

Exemple : Soit $X_1, \ldots, X_n$ i.i.d. $\sim \mathcal{N}(\mu, \sigma^2)$. Considérons la statistique

$$T_n = T(X_1, \ldots, X_n) = \frac{1}{n-1} \sum_{j=1}^{n} (X_i - \bar{X}_n)^2.$$

On définit

$$Y_n = (n-1)\frac{T_n}{\sigma^2}.$$

D'après le théorème de Fisher (cf. Chapitre 1), la v.a. Y_n est distribuée selon χ^2_{n-1}, la loi Khi-deux à $n-1$ degrés de liberté. La densité de Y_n est donnée par

$$f_{Y_n}(y) = \frac{1}{2^{\frac{n-1}{2}}\Gamma(\frac{n-1}{2})}\ y^{\frac{n-3}{2}}\ e^{-y/2} 1_{y \geq 0}.$$

En effectuant le changement de variable $Y_n \to T_n$, on obtient la densité de T_n :

$$f_{T_n}(t,\sigma^2) = \frac{1}{2^{\frac{n-1}{2}}\Gamma(\frac{n-1}{2})} \left(\frac{n-1}{\sigma^2}\right)^{\frac{n-1}{2}} t^{\frac{n-3}{2}} \; e^{-\frac{(n-1)t}{2\sigma^2}} 1_{t\geq 0}.$$

Calcul de l'information de Fisher $I_{T_n}(\sigma^2)$:

$$\log f_{T_n}(t,\sigma^2) = constante(t) - \frac{(n-1)t}{2\sigma^2} + \frac{n-1}{2}\log\left(\frac{n-1}{\sigma^2}\right)$$

où $constante(t)$ ne dépend pas de σ.

$$\frac{\partial^2 \log f_{T_n}(t,\sigma^2)}{\partial(\sigma^2)^2} = -\frac{(n-1)t}{\sigma^6} + \frac{n-1}{2\sigma^4}$$

d'où

$$I_{T_n}(\sigma^2) = \frac{n-1}{\sigma^6}E(T_n) - \frac{n-1}{2\sigma^4} = \frac{n-1}{\sigma^4} - \frac{n-1}{2\sigma^4} = \frac{n-1}{2\sigma^4}.$$

D'autre part, on sait que l'information de Fisher sur σ^2 contenue dans l'échantillon $X_1,\ldots,X_n$ vaut

$$I_n(\sigma^2) = nI(\sigma^2) = \frac{n}{2\sigma^4} \quad \text{(par additivité de l'information).}$$

Il s'en suit que pour une taille d'échantillon finie n, la variance empirique (modifiée), T_n, n'est pas exhaustive pour σ^2 puisque $I_{T_n}(\sigma^2) < I_n(\sigma^2)$.

3.4 Cas des familles exponentielles

La plupart des lois usuelles font partie de ce qu'on appelle la famille exponentielle.

Définition 3.4.1 *Une loi de probabilité $P_\theta, \theta \in \Theta$, de densité $f(x,\theta)$ relativement à une mesure σ-finie ν est dite appartenir à une famille exponentielle s'il existe des fonctions $\theta \mapsto \alpha_j(\theta), \theta \mapsto c(\theta)$, $x \mapsto T_j(x)$ et $x \mapsto h(x)$ $(h(x) > 0)$, telles que*

$$f(x,\theta) = c(\theta)h(x)\exp\left(\sum_{j=1}^{r}\alpha_j(\theta)T_j(x)\right).$$

Exemples : Soit $X \sim \mathcal{N}(\mu, \sigma^2)$, $\mu \in \mathbb{R}$ et $\sigma^2 > 0$.

$$\begin{aligned} f(x,\mu,\sigma) &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \\ &= \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma} \exp\left(-\frac{\mu^2}{2\sigma^2}\right) \exp\left(-\frac{x^2}{2\sigma^2} + \frac{\mu}{\sigma^2} x\right) \end{aligned}$$

$$\begin{aligned} c(\theta) &= \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma} \exp\left(-\frac{\mu^2}{2\sigma^2}\right) \\ h(x) &= 1 \\ \alpha_1(\theta) &= -\frac{1}{2\sigma^2} \\ \alpha_2(\theta) &= \frac{\mu}{\sigma^2} \\ T_1(x) &= x^2 \\ T_2(x) &= x \end{aligned}$$

Posons $\theta_j = \alpha_j(\theta)$.

Définition 3.4.2 *Les paramètres (θ_j) s'appellent les paramètres naturels de la famille exponentielle*

$$f(x,\theta) = b(\theta) h(x) \exp\left(\sum_{j=1}^{r} \theta_j T_j(x)\right).$$

Exemples :

- Soit $X \sim \mathcal{N}(\mu, \sigma^2)$, $\mu \in \mathbb{R}$ et $\sigma^2 > 0$. On déduit de l'exemple précédent que la paramétrisation naturelle de la loi normale de moyenne μ et de variance σ^2 est donnée par

  $$f(x,\theta_1,\theta_2) = \frac{1}{\sqrt{\pi}} \sqrt{-\theta_1} \exp\left(-\frac{\theta_2^2}{4\theta_1}\right) \exp(\theta_1 x^2 + \theta_2 x)$$

  où

  $$\begin{aligned} \theta_1 &= -\frac{1}{2\sigma^2} \\ \theta_2 &= \frac{\mu}{\sigma^2} \end{aligned}$$

- Soit $X \sim B(n,p)$, $0 < p < 1$. $f(x,p) = C_n^x p^x (1-p)^{n-x}$ qui s'écrit aussi

  $$f(x,p) = C_n^x (1-p)^n \exp\left(x \ \log \frac{p}{1-p}\right).$$

d'où, en posant $\theta = \log(p/(1-p))$,

$$\begin{aligned} b(\theta) &= \frac{1}{(1+\exp(\theta))^n} \\ h(x) &= C_n^x \\ T_1(x) &= T(x) = x. \end{aligned}$$

– Soit $X \sim \gamma(\alpha, \beta)$ avec $\alpha, \beta > 0$. $f(x, \alpha, \beta) = \beta^\alpha [\Gamma(\alpha)]^{-1} x^{\alpha-1} \exp(-\beta\, x),\ x > 0$.

$$f(x, \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \exp\left[-\beta\, x + \log(x)(\alpha - 1)\right]$$

Si on pose

$$\theta_1 = \beta, \quad \theta_2 = \alpha - 1$$

alors la paramétrisation naturelle de la loi gamma est donnée par

$$\begin{aligned} b(\theta_1, \theta_2) &= \frac{\theta_1^{\theta_2+1}}{\Gamma(\theta_2+1)} \\ h(x) &= 1 \\ T_1(x) &= -x \\ T_2(x) &= \log(x). \end{aligned}$$

La famille exponentielle admet de nombreuses propriétés. En particulier, on admettra le résultat suivant qui nous donne un exemple de statistique exhaustive complète.

Théorème 3.4.1 *Soient $\Theta \subset \mathbb{R}^d$ l'espace des paramètres et $(X_1, \ldots, X_n)$ un échantillon de $X \sim P_\theta$ appartenant à une famille exponentielle de densité*

$$f_X(x, \theta) = c(\theta) h(x) \exp\left(\sum_{j=1}^{r} \alpha_j(\theta) T_j(x)\right)$$

tel que $d \leq r$, les fonctions $T_1, ..., T_r$ (définies sur $\mathcal{X}$) sont affinement indépendantes ainsi que les fonctions $\alpha_1, .., \alpha_r$ (définies sur Θ) et

$$\Big\{(\alpha_1(\theta), ..., \alpha_r(\theta)),\ \theta \in \Theta\Big\} \text{ contient un ouvert de dimension égale à } r.$$

Alors,

$$\left(\sum_{j=1}^{n} T_1(X_j), \ldots, \sum_{j=1}^{n} T_r(X_j)\right)$$

est une statistique exhaustive complète pour $\theta = (\theta_1, \ldots, \theta_d)$.

Remarque. Des fonctions $f_1, ..., f_k$ sont dites affinement indépendantes si

$$c_1 f_1 + ...c_k f_k = c_{k+1} \implies c_1 = ... = c_k = c_{k+1} = 0.$$

Théorème 3.4.2 *Si la paramétrisation naturelle du modèle appartenant à la famille exponentielle est donnée par*

$$f(x, \theta) = b(\theta)h(x) \exp\left(\sum_{j=1}^{d} \theta_j T_j\right), \theta \in \Theta$$

où Θ est un ouvert de $\mathbb{R}^d$, alors une condition nécessaire et suffisante pour que le modèle soit identifiable est que $I(\theta)$ soit inversible.

Deuxième partie

L'estimation statistique

Chapitre 4

Généralités

4.1 Introduction

Soit X une v.a. à valeurs dans $(\mathcal{X}, \mathcal{B})$ de loi $P_\theta, \theta \in \Theta$. La densité de P_θ par rapport à une mesure dominante σ- finie ν (mesure de comptage dans le cas d'une loi discrète, la mesure de Lebesgue dans le cas d'une loi absolument continue) sera notée $f(x, \theta)$.

L'objectif du statisticien est de connaître la vraie valeur du paramètre θ, ou plus généralement une fonction de cette valeur, $g(\theta)$ où g est une application $\Theta \to \mathbb{R}^d$. Grâce à l'information fournie par un échantillon $X_1, \ldots, X_n$ iid de la loi P_θ, le statisticien tentera d'approximer $g(\theta)$ en choisissant ce qu'on appelle un estimateur, c'est à dire une statistique à valeur dans $g(\Theta)$. Un estimateur est une fonction mesurable T de l'échantillon aléatoire $(X_1, \ldots, X_n)$, lui même variable aléatoire de $g(\Theta)$ noté $T(X_1, \ldots, X_n)$.

Lorsque g est égale à l'identité, on notera parfois $\hat{\theta}_n$ la valeur de l'estimateur $T(X_1, \ldots, X_n)$ de θ. En général, on utilisera la notation T_n pour l'estimateur de $g(\theta)$ lorsque g est une application g quelconque. On utilisera parfois d'autres notations qui sont traditionnellement réservées à certains estimateurs/statistiques ($\bar{X}_n$ pour la moyenne empirique, S_n^2 pour la variance empirique...).

Dans ce qui suit, on utilisera la notation H_{θ^2} pour l'opérateur Hessien.

4.2 Propriétés générales d'un estimateur

Soit T_n un estimateur de $g(\theta)$. On rappelle que la notation T_n désigne la variable $T(X_1, \ldots, X_n)$ où $(X_1, \ldots, X_n)$ est un échantillon aléatoire de l'espace $(\mathcal{X}^n, P_\theta), \theta \in \Theta$. On considère ici que $g(\Theta) \subset \mathbb{R}^d$ et on munit $\mathbb{R}^d$ de la norme $\|\|$ associée au produit scalaire usuel.

4.2.1 Estimateur sans biais

Définition 4.2.1 *La quantité $b_n(\theta) = \mathbb{E}_\theta(T_n) - g(\theta)$ s'appelle biais de l'estimateur T_n. Un estimateur T_n de $g(\theta)$ est dit sans biais ou non-biaisé si*

$$b_n(\theta) = 0 \quad soit \quad \mathbb{E}(T_n) = g(\theta).$$

Exemple : Soit le modèle (X, P_θ) tel que $\mathbb{E}(X) = \mu$ et $Var(X) = \sigma^2$ existent. La moyenne empirique et la variance empirique modifiée

$$\frac{1}{n-1}\sum_{j=1}^{n}(X_j - \bar{X}_n)^2$$

sont respectivement des estimateurs non-biaisés de la vraie espérance μ et la vraie variance σ^2.

Remarque : La définition du biais nécessite l'intégrabilité de T_n.

4.2.2 Estimateur asymptotiquement sans biais

Définition 4.2.2 *Un estimateur T_n de $g(\theta)$ est dit asymptotiquement sans biais si*

$$\lim_{n\to\infty} b_n(\theta) = 0 \quad soit \quad \lim_{n\to\infty} \mathbb{E}(T_n) = g(\theta).$$

Exemple : Supposons toujours que $Var(X) = \sigma^2$ est finie. La variance empirique

$$S_n^2 = \frac{1}{n}\sum_{j=1}^{n}(X_j - \bar{X}_n)^2$$

est un estimateur de σ^2 qui est asymptotiquement sans biais :

$$\mathbb{E}(S_n^2) = \frac{n-1}{n}\,\sigma^2 \to \sigma^2 \ \text{ lorsque } \ n \to \infty.$$

4.2.3 Estimateur convergent

Définition 4.2.3 *T_n est un estimateur convergent s'il converge en probabilité vers $g(\theta)$*

$$\lim_{n\to\infty} P(\|T_n - g(\theta)\| > \epsilon) = 0, \quad \forall\ \epsilon > 0.$$

On notera $T_n \xrightarrow{P} g(\theta)$.

Théorème 4.2.1 *Un estimateur T_n asymptotiquement sans biais tel que $\lim_{n\to\infty} \mathbb{E}\|T_n - \mathbb{E}(T_n)\|^2 = 0$ est convergeant.*

(cf. Chapitre 1 pour les propriétés des modes de convergence).

4.3 Comparaison des estimateurs

Soient T_n et T'_n deux estimateurs de $g(\theta)$ de biais $b_n(\theta) = \mathbb{E}(T_n) - g(\theta)$ et $b'_n(\theta) = \mathbb{E}(T'_n) - g(\theta)$. Comment faut-il comparer T_n et T'_n ? On doit donc choisir un critère qui permettra au statisticien de mesurer l'efficacité de chacun des estimateurs. Un bon critère est le risque quadratique :

$$\mathbb{E}\|T_n - g(\theta)\|^2.$$

Remarque : Le risque quadratique est un critère d'efficacité des estimateurs T_n de carrés intégrables.

4.3.1 Décomposition biais-variance du risque

On a la décomposition biais-variance du risque quadratique :

$$\begin{aligned}
\mathbb{E}[\|T_n - g(\theta)\|^2] &= \mathbb{E}[\|T_n - g(\theta) - b_n(\theta) + b_n(\theta)\|^2] \\
&= \mathbb{E}[\|T_n - g(\theta) - b_n(\theta)\|^2] + b_n^2(\theta) \\
&= \mathbb{E}(T_n - g(\theta) - b_n(\theta))^T(T_n - g(\theta) - b_n(\theta)) + b_n^2(\theta) \\
&= Tr(Var(T_n - g(\theta))) + b_n^2(\theta) = Tr(Var(T_n)) + b_n^2(\theta).
\end{aligned}$$

où $Var(T_n)$ est la matrice variance-covariance de T_n.

4.3.2 Comparaison des variances des estimateurs sans biais

La comparaison d'estimateurs sans biais revient à comparer leurs variances, d'où

Définition 4.3.1 *Soient T_n et T'_n deux estimateurs sans biais de $g(\theta)$. T'_n est dit plus efficace que T_n s'il est préférable au sens de la variance :*

$$Var_\theta(T'_n) \leq Var_\theta(T_n) \quad \forall\ \theta \in \Theta.$$

On dit que l'estimateur sans biais T'_n est uniformément plus efficace si il est plus efficace que tous les estimateurs sans biais. On dit aussi qu'il est de variance minimale.

On rappelle que pour deux matrices A et B on a $A \leq B \Leftrightarrow B - A$ est une matrice symétrique positive. La notation Var_θ marque bien la dépendance de la variance du modèle P_θ et donc du paramètre inconnu $\theta \in \Theta$. Le critère d'efficacité n'a de sens que pour discriminer les estimateur sans biais.

Théorème 4.3.1 (Lehmann-Scheffé) *Si T_n est un estimateur sans biais de $g(\theta)$ et si S_n est une statistique exhaustive et complète, alors l'unique estimateur de $g(\theta)$ sans biais uniformément de variance minimale est $T'_n = \mathbb{E}_\theta(T_n \mid S_n)$.*

Notons que le théorème précédent implique que l'estimateur T_n' est une fonction de S_n. Malheureusement les statistiques exhaustives complètes n'existent pas toujours. On recherche un critère absolu, à savoir s'il existe une borne inférieure non triviale à l'ensemble des variances des estimateurs T_n sans biais de $g(\theta)$. On cherche donc

$$\min_{T_n \in B_0} Var(T_n)$$

où B_0 est l'ensemble des estimateurs sans biais de $g(\theta)$.

Soit $T_n = T(X_1, ..., X_n)$ un estimateur sans biais de $g(\theta), \theta \in \Theta$.

Hypothèses : (les hypothèses H1-H4 sont les hypothèses usuelles faites au Chapitre 3)

H1 : Θ est un ouvert de $\mathbb{R}^p$ pour $p \geq 1$.

H2 : Le support $\{x : f(x, \theta) > 0\}$ ne dépend pas de θ.

H3 : Pour tout x la fonction $\theta \mapsto f(x, \theta)$ est au moins deux fois continuement-dérivable sur Θ.

H4 : Pour tout $B \in \mathcal{B}$ l'intégrale $\int_B f(x, \theta) d\nu(x)$ est au moins deux fois dérivable sous le signe d'intégration et on peut permuter intégration et dérivation.

H5' : La statistique $T_n = T(X_1, \ldots, X_n)$ est de carré intégrable : elle satisfait $\mathbb{E}(T_n^2) < +\infty$ et

$$\begin{aligned}
&\frac{\partial}{\partial \theta} E_\theta(T_n) \\
&= \frac{\partial}{\partial \theta} \int_{\mathcal{X}^n} T(x_1, ..., x_n) f(x_1, ..., x_n) d\nu(x_1) ... d\nu(x_n) \\
&= \int_{\mathcal{X}^n} T(x_1, ..., x_n) \frac{\partial}{\partial \theta} f(x_1, ..., x_n) d\nu(x_1) ... d\nu(x_n)
\end{aligned}$$

H6' : La fonction g est dérivable sur Θ. On note $J_\theta(g)$ sa matrice Jacobienne de taille $d \times p$ en tout point $\theta \in \Theta$.

Théorème 4.3.2 *Supposons que les hypothèses H1-H4, H5' et H6' sont vérifiées. Si $T_n = T(X_1, \ldots, X_n)$ est un estimateur sans biais de $g(\theta)$ alors*

$$Var(T_n) \geq \nabla_\theta g(\theta) I_n^{-1}(\theta) (\nabla_\theta g(\theta))^T.$$

La quantité $\nabla_\theta g(\theta) I_n^{-1}(\theta) (\nabla_\theta g(\theta))^T$ est appelée borne de Cramér-Rao.

Remarques :

– L'hypothèse que $\int_B f(x,\theta)d\nu(x)$ est dérivable deux fois sous le signe d'intégration n'est pas réellement nécessaire pour établir l'inégalité. Lorsqu'elle est vérifiée, on sait qu'alors

$$I(\theta) = -E\left[H_{\theta^2}(\log f)(X,\theta)\right].$$

– Lorsque g est égale à l'identité, c-à-d $g(\theta) = \theta$ alors l'inégalité se réduit à

$$Var(T_n) \geq I_n^{-1}(\theta) = \frac{1}{n} I^{-1}(\theta).$$

– Rien ne garantie l'existence d'un estimateur dont la variance atteint la borne de Cramér-Rao.

4.3.3 Efficacité d'un estimateur

Définition 4.3.2 *Un estimateur T_n sans biais pour $g(\theta)$ est dit efficace si sa variance atteint la borne de Cramér-Rao.*

Définition 4.3.3 *Un estimateur T_n sans biais de $g(\theta)$ est asymptotiquement efficace si*

$$\lim_{n\to\infty} Var(T_n)(\nabla_\theta g(\theta) I_n^{-1}(\theta)(\nabla_\theta g(\theta))^T)^{-1} = 1.$$

Remarques :

– Un estimateur efficace est de variance minimale .

– Un estimateur peut être sans biais, de variance minimale, mais ne pas atteindre la borne de Cramér-Rao, donc ne pas être efficace. Dans ce cas-là, la borne Cramér-Rao est "trop petite" pour être atteinte. Voir l'exemple ci-dessous du modèle de Poisson.

Exemples :

- Soit $X \sim \mathcal{N}(\mu, \sigma^2)$.
  – On a (cf. Chapitre 3)

$$I_n(\mu) = \frac{n}{\sigma^2}.$$

D'autre part (cf. Chapitre 2)

$$Var(\bar{X}_n) = \frac{\sigma^2}{n}.$$

Il s'en suit que la moyenne empirique est un estimateur efficace pour μ.

– La variance modifiée

$$\frac{1}{n-1}\sum_{j=1}^{n}(X_j-\bar{X}_n)^2$$

est un estimateur asymptotiquement efficace pour σ^2.

• Soit $X\sim\mathcal{P}(\theta),\theta>0$. Considérons l'estimation de $g(\theta)=\theta^2$. On peut facilement montrer que $\bar{X}_n$ est exhaustive complète. D'autre part, l'estimateur

$$T_n=(\bar{X}_n)^2-\frac{1}{n}\bar{X}_n$$

est un estimateur sans biais de $g(\theta)$. En effet,

$$\begin{aligned}E_\theta(T_n) &= Var_\theta(\bar{X}_n)+[E_\theta(\bar{X}_n)]^2-\frac{1}{n}E_\theta(\bar{X}_n)\\ &= \frac{1}{n}\theta+\theta^2-\frac{1}{n}\theta\\ &= g(\theta).\end{aligned}$$

Par le Théorème de Lehmann-Scheffé, T_n est l'estimateur sans biais uniformément de variance minimale (T_n est une fonction de la statistique exhaustive complète $\bar{X}_n$). Cependant, l'estimateur T_n n'est pas efficace. En effet, les hypotèses $H1-H5$ sont vérifiées puisque

– $H1:\Theta=]0,\infty[$ est ouvert,

– $H2-H4$: La densité du modèle

$$f(x,\theta)=\frac{1}{x!}e^{-\theta}\theta^x=\frac{1}{x!}e^{-\theta}\exp\{x\log(\theta)\}$$

appartient à une famille exponentielle avec $\alpha(\theta)=\log(\theta)$ une fonction $C^2(\Theta)$.

– $H5:Var(T_n)<\infty$ est finie puisque $Var(X_i)<\infty$. D'autre part,

$$\begin{aligned}E_\theta(T_n) &= \int_{\mathbb{N}^n}\left((\bar{x}_n)^2-n^{-1}\bar{x}_n\right)f(x_1,...,x_n,\theta)d\nu(x_1)...d\nu(x_n),\\ &\quad (\nu:\text{ la mesure de comptage})\\ &= \frac{e^{-n\theta}}{n^2}\sum_{x_n=0}^{\infty}\cdots\sum_{x_1=0}^{\infty}\left((x_1+...+x_n)^2-(x_1+...+x_n)\right)\frac{\theta^{x_1+...+x_n}}{x_1!\ldots x_n!}.\end{aligned}$$

Le second terme de l'expression précédente (la somme) est une série entière de rayon de convergence égal à ∞ et donc elle est dérivable terme à terme

(sur $\mathbb{C}$ et en particulier sur Θ), c'est-à-dire pour tout $\theta > 0$ on a que

$$= \frac{\partial}{\partial\theta} \sum_{x_n=0}^{\infty} \dots \sum_{x_1=0}^{\infty} \left((x_1 + ... + x_n)^2 - (x_1 + ... + x_n)\right) \frac{\theta^{x_1+...+x_n}}{x_1! \dots x_n!}$$
$$= \sum_{x_n=0}^{\infty} \dots \sum_{x_1=0}^{\infty} (x_1 + ... + x_n) \left((x_1 + ... + x_n)^2 - (x_1 + ... + x_n)\right) \frac{\theta^{x_1+...+x_n-1}}{x_1! \dots x_n!}$$

et donc

$$\frac{\partial}{\partial\theta} E_\theta(T_n)$$
$$= -\frac{e^{-n\theta}}{n} \sum_{x_n=0}^{\infty} \dots \sum_{x_1=0}^{\infty} \left((x_1 + ... + x_n)^2 - (x_1 + ... + x_n)\right) \frac{\theta^{x_1+...+x_n}}{x_1! \dots x_n!}$$
$$+\frac{e^{-n\theta}}{n^2} \sum_{x_n=0}^{\infty} \dots \sum_{x_1=0}^{\infty} (x_1 + ... + x_n) \left((x_1 + ... + x_n)^2 - (x_1 + ... + x_n)\right) \frac{\theta^{x_1+...+x_n-1}}{x_1! \dots x_n!}$$
$$= \int_{\mathbb{N}^n} \left((\bar{x}_n)^2 - n^{-1}\bar{x}_n\right) \frac{\partial}{\partial\theta} f(x_1, ..., x_n, \theta) d\nu(x_1) ... d\nu(x_n).$$

D'autre part,

$$\begin{aligned} \frac{\partial^2}{\partial\theta^2} \log(f(x,\theta)) &= \frac{\partial^2}{\partial\theta^2}(-\theta + x\log(\theta)) \\ &= -\frac{x}{\theta^2} \end{aligned}$$

d'où

$$I_n(\theta) = n\frac{E_\theta(X)}{\theta^2} = \frac{n}{\theta}.$$

La borne de Cramér-Rao est donnée par

$$\frac{[g'(\theta)]^2}{I(\theta)} = \frac{4\theta^3}{n}.$$

Enfin, on calcule la variance de T_n :

$$\begin{aligned} Var(T_n) &= Var((\bar{X}_n)^2 - n^{-1}\bar{X}_n) \\ &= E\left[\left((\bar{X}_n)^2 - n^{-1}\bar{X}_n\right)^2\right] - \theta^4 \\ &= n^{-4}E\left[(X_1 + ... + X_n)^4 - 2(X_1 + ... + X_n)^3 + (X_1 + ... + X_n)^2\right] - \theta^4 \end{aligned}$$

où $X_1 + ... + X_n \sim \mathcal{P}(n\theta)$.

$E((X_1 + ... + X_n)^2) = Var(X_1 + ... + X_n) + (E(X_1 + ... + X_n))^2 = n\theta + n^2\theta^2,$

$$\begin{aligned}
E((X_1+...+X_n)^3) &= e^{-n\theta}\sum_{k=0}^{\infty}\frac{k^3(n\theta)^k}{k!}\\
&= e^{-n\theta}\sum_{k=1}^{\infty}\frac{k^2(n\theta)^k}{(k-1)!}\\
&= e^{-n\theta}n\theta\sum_{k=0}^{\infty}\frac{(k+1)^2(n\theta)^k}{k!}\\
&= n\theta\left(\sum_{k=0}^{\infty}e^{-n\theta}\frac{k^2(n\theta)^k}{k!}+2\sum_{k=0}^{\infty}e^{-n\theta}\frac{k(n\theta)^k}{k!}+\sum_{k=0}^{\infty}e^{-n\theta}\frac{(n\theta)^k}{k!}\right)\\
&= n\theta\left(n\theta+n^2\theta^2+2n\theta+1\right)\\
&= n\theta\left(n^2\theta^2+3n\theta+1\right).
\end{aligned}$$

De la même manière, on montre que

$$E((X_1+...+X_n)^4) = n\theta\left(n\theta(n^2\theta^2+3n\theta+1)+3(n\theta+n^2\theta^2)+3n\theta+1\right).$$

Il vient que

$$\begin{aligned}
Var(T_n) &= n^{-4}n\theta\left(n^3\theta^3+4n^2\theta^2+2n\theta\right))-\theta^4\\
&= \frac{4\theta^3}{n}+\frac{2\theta^2}{n^2}>\frac{4\theta^3}{n}
\end{aligned}$$

et donc T_n n'est pas efficace.

Les modèles exponentiels jouent un rôle central pour l'estimation paramétrique puisque sous certaines conditions ils sont les seuls pour lesquels on aura une estimation sans biais efficace. Cela représente une restriction du critère sans biais et efficace.

Théorème 4.3.3 *Soit un modèle statistique paramétrique dominé et régulier. Soit $g:\mathbb{R}^d\to\mathbb{R}^d$ de classe $\mathcal{C}^1$ sur Θ telle que la matrice Jacobienne carrée de taille d, $J_\theta(g)$ soit inversible pour tout $\theta\in\Theta$. Alors T_n est un estimateur sans biais de $g(\theta)$ atteignant la borne de Cramer-Rao si et seulement si le modèle est exponentiel.*

Chapitre 5

Méthodes d'estimation ponctuelle

Nous présenterons dans la suite trois méthodes d'estimation. Nous verrons celle du maximum de vraisemblance et celle des moments qui relèvent de l'approche classique. Ensuite, nous donnerons quelques notions sur les estimateurs Bayésiens.

5.1 Maximum de vraisemblance

5.1.1 Définition et caractéristiques

Définition 5.1.1 *Soit le modèle statistique $(\mathcal{X}, P_\theta)$ et $f(x,\theta)$ la densité de P_θ relativement à la mesure dominante ν (mesure de comptage dans le cas des lois discrètes, mesure de Lebesgue dans le cas des lois absolument continues). Considérons un échantillon aléatoire $(X_1, \ldots, X_n)$ de $(\mathcal{X}, P_\theta)$. On appelle vraisemblance la variable aléatoire*

$$L_n(\theta) = f(X_1, \ldots, X_n, \theta).$$

es variables $X_j, j = 1, \ldots, n$ étant iid, on a

$$f(X_1, \ldots, X_n, \theta) = \prod_{j=1}^{n} f(X_j, \theta).$$

Définition 5.1.2 *Soit $L_n(\theta)$ la vraisemblance au point $\theta \in \Theta$. On appelle estimateur du Maximum de Vraisemblance (EMV) la statistique $\hat{\theta}_n = \hat{\theta}(X_1, \ldots, X_n)$ telle que*

$$L_n(\hat{\theta}_n) \geq L_n(\theta) \quad \forall\ \theta \in \Theta \qquad p.s.$$

Caractéristique 1. L'EMV n'existe pas toujours.
En effet la maximisation se fait sur un ensemble ouvert Θ. Par contre pour tout voisinage compact $V(\theta)$ de tout point $\theta \in \Theta$ tel que $V(\theta) \subseteq \Theta$ un maximum local $\hat{\theta}_n$ existe dès que L_n est continue. On parle alors de maximum de vraisemblance local.

Caractéristique 2. La vraisemblance n'est pas a priori dérivable en tout point $\theta \in \Theta$.

Exemple : Soient $X_1, \ldots, X_n \sim U[0, \theta]$.

$$\begin{aligned} L_n(\theta) = \prod_{j=1}^{n} \frac{1}{\theta} 1_{[0,\theta]}(X_j) &= \frac{1}{\theta^n} 1_{0 \,\leq\, \inf_{1\leq j\leq n} X_j \,\leq\, \sup_{1\leq j\leq n} X_j \,\leq\, \theta} \\ &= \frac{1}{\theta^n} 1_{0 \,\leq\, \inf_{1\leq j\leq n} X_j} \times 1_{\theta \,\geq\, \sup_{1\leq j\leq n} X_j} \end{aligned}$$

et donc

$$\hat{\theta}_n = \sup_{1\leq j\leq n} X_j$$

est le EMV.

Caractéristique 3. Il n'y aucune raison pour que le EMV soit sans biais.

Exemple : Reprenons l'exemple précédent. Posons $Y_n = \sup_{1\leq j\leq n} X_j$. On peut montrer que (exercice)

$$\theta Y_n \sim Beta(n, 1);$$

i.e., la loi de la variable aléatoire Y_n admet pour densité

$$f(y, \theta) = \frac{n y^{n-1}}{\theta} 1_{0\leq y\leq \theta}.$$

Il s'en suit que

$$E(Y_n) = \frac{n}{n+1}\ \theta$$

et

$$b_n(\theta) = E(Y_n) - \theta = -\frac{1}{n+1}\ \theta \neq 0.$$

Caractéristique 4. L'EMV n'a aucune raison d'être unique.

Exercice : Soient $X_1, \cdots, X_n \sim U[\theta, \theta+1]$. Montrer que tout estimateur $\hat{\theta}_n$ de θ compris entre $\sup_{1\leq i\leq n} X_i - 1$ et $\inf_{1\leq i\leq n} X_i$ est un estimateur de MV.

Dans toute la suite, on suppose que les hypothèses usuelles du Chapitre 3 sont vérifiées ainsi que l'hypothèse d'identifiabilité :

H0 : $\theta \neq \theta' \Rightarrow P_\theta \neq P_{\theta'}$.

H1 : Θ est un ouvert de $\mathbb{R}^d$.

H2 : Le support $\{x : f(x,\theta) > 0\}$ ne dépend pas de θ.

H3 : Pour tout x la fonction $\theta \mapsto f(x,\theta)$ est au moins deux fois continuement-dérivable sur Θ.

H4 : Pour tout $B \in \mathcal{B}$ l'intégrale $\int_B f(x,\theta)d\nu(x)$ est au moins deux fois dérivable sous le signe d'intégration et on peut permuter intégration et dérivation.

L'estimateur de Maximum de Vraisemblance, $\hat{\theta}_n$, est solution (P_θ-p.s.) du système

$$\begin{cases} \nabla_{\hat{\theta}_n} L_n = 0 \\ \\ H_{\hat{\theta}_n^2} L_n < 0. \end{cases} \tag{5.1}$$

Il s'avère qu'il est plus facile d'utiliser le logarithme de la vraisemblance $l_n(\theta)$:

$$\begin{aligned} l_n(\theta) &= \log L_n(\theta) \\ &= \sum_{j=1}^{n} \log f(X_j, \theta). \end{aligned}$$

Comme le logarithme est une fonction croissante, $\hat{\theta}_n$ est aussi le maximum de $l_n(\theta), \theta \in \Theta$. Le système (5.1) est équivalent au système (P_θ-p.s.)

$$\begin{cases} \sum_{j=1}^n \nabla_{\hat{\theta}_n} \log f(X_j, \cdot) = 0 \\ \\ \sum_{j=1}^n H_{\hat{\theta}_n^2} \log f(X_j, \cdot) < 0. \end{cases} \tag{5.2}$$

La première équation du système (5.2) est appelée l'équation de vraisemblance ou condition du premier ordre. La seconde est la condition du second ordre.

Remarques :

- La condition du premier ordre s'écrit aussi $\sum_{j=1}^n S(X_j, \hat{\theta}_n) = 0$ p.s. où S est la fonction score (cf. Chapitre 3).
- Si $\sum_{j=1}^n H_{\theta^2} \log f(X_j, \cdot) < 0$ pour tout $\theta \in \Theta$, alors L_n est strictement concave et une solution $\hat{\theta}_n$ de la condition du premier ordre est unique. Si elle existe, c'est l'EMV.

5.1.2 Quelques exemples

• **Calcul de l'EMV du paramètre d'une loi $\mathcal{P}(\lambda)$**

Soient $X_1, \ldots, X_n \sim \mathcal{P}(\lambda)$. On a

$$L_n(\lambda) = e^{-n\lambda}\lambda^{\sum_{j=1}^n X_j} \prod_{j=1}^n \frac{1}{X_j!}$$

$$l_n(\lambda) = -n\lambda + \sum_{j=1}^n X_j \log\lambda + cte.$$

$$\frac{\partial l_n(\hat{\lambda}_n)}{\partial\lambda} = -n + \frac{\sum_{j=1}^n X_j}{\hat{\lambda}_n} = 0 \Leftrightarrow \hat{\lambda}_n = \bar{X}_n$$

et

$$\frac{\partial^2 l_n(\hat{\lambda}_n)}{\partial^2\lambda} = -\frac{\sum_{j=1}^n X_j}{\hat{\lambda}_n^2} < 0.$$

Donc, $\hat{\lambda}_n = \bar{X}_n$ (la moyenne empirique) est l'EMV dans le cas d'un modèle de Poisson.

• **Calcul de l'EMV de l'espérance d'une loi $\mathcal{N}(\mu, \sigma^2)$, $\sigma = \sigma_0$ connu**
On a la vraisemblance

$$L_n(\mu) = \frac{1}{(2\pi)^{n/2}\sigma_0^n} e^{-\sum_{j=1}^n \frac{(X_j-\mu)^2}{2\sigma_0^2}}$$

$$l_n(\mu) = cte - \sum_{j=1}^n \frac{(X_j-\mu)^2}{2\sigma_0^2}$$

$$\frac{\partial l_n(\hat{\mu}_n)}{\partial\mu} = \frac{1}{\sigma_0^2}\sum_{j=1}^n (X_j - \mu) = 0 \Leftrightarrow \hat{\mu}_n = \bar{X}_n$$

et

$$\frac{\partial^2 l_n(\mu)}{\partial^2\mu}|_{\mu=\hat{\mu}_n} = -\frac{n}{\sigma_0^2} < 0.$$

L'EMV de l'espérance d'une loi normale est égal à la moyenne empirique que l'écart-type, σ, soit connu ou inconnu.

• **Calcul de l'EMV du paramètre d'une famille exponentielle**

Considérons une famille exponentielle sous sa forme naturelle

$$f(x,\lambda) = b(\lambda)h(x)\exp\left(\sum_{i=1}^{d}\lambda_i T_i(x)\right) = h(x)\exp\left(\sum_{i=1}^{d}\lambda_i T_i(x) + \beta(\lambda)\right), \quad \lambda \in \Lambda$$

où $\beta(\lambda) = \log b(\lambda)$. Si
- $\beta(\lambda)$ est 2 fois continûment dérivable
- $H_{\lambda^2}(\beta)(\lambda) < 0, \ \forall\ \lambda \in \Lambda$

alors $\hat{\lambda}_n$ est l'EMV de $\lambda = (\lambda_1, \ldots, \lambda_d)'$ si et seulement si

$$\frac{\partial\beta(\hat{\lambda}_n)}{\partial\lambda_i} = -\frac{1}{n}\sum_{j=1}^{n} T_i(X_j) \quad \text{pour tout } i.$$

Dans ce cas, L'EMV $\hat{\lambda}_n$ existe et il est unique. En effet,

$$\begin{aligned}
L_n(\lambda) &= \prod_{j=1}^{n} h(X_j)\exp\left(\sum_{j=1}^{n}\sum_{i=1}^{d}\lambda_i T_i(X_j) + n\beta(\lambda)\right) \\
l_n(\lambda) &= cte + \sum_{j=1}^{n}\sum_{i=1}^{d}\lambda_i T_i(X_j) + n\beta(\lambda) \\
\nabla_\lambda l_n(\hat{\lambda}_n) &= \frac{\partial\beta(\hat{\lambda}_n)}{\partial\lambda_i} = -\frac{1}{n}\sum_{j=1}^{n} T_i(X_j) \quad \text{pour tout } i, \\
H_{\lambda^2}(l_n)(\hat{\lambda}_n) &= nH_{\lambda^2}(\beta)(\hat{\lambda}_n) < 0.
\end{aligned}$$

En plus, les hypothèses garantissent que l'équation de vraisemblance admette une solution unique (notons que ces hypothèses impliquent que la fonction $\beta(\lambda)$ est concave).

5.1.3 Propriétés à distance finie de l'EMV

On s'intéresse ici au cas où n est fixé. On suppose qu'il existe un unique EMV $\hat{\theta}_n$ de θ.

Théorème 5.1.1 *Si T_n est une statistique exhaustive pour θ alors $\hat{\theta}_n$ est une fonction de T_n.*

Démonstration : D'après le critère de factorisation, on peut trouver deux fonctions positives h et g telles que

$$L_n(\theta) = f(X_1, \ldots, X_n, \theta) = h(X_1, \ldots, X_n) g(T_n, \theta)$$

et donc

$$\max_{\theta \in \Theta} L_n(\theta) = \max_{\theta \in \Theta} h(T_n, \theta)$$

et par conséquent l'EMV $\hat{\theta}_n$, qui satisfait par définition

$$h(T_n, \hat{\theta}_n) \geq h(T_n, \theta), \quad \forall \theta \in \Theta,$$

s'écrit sous la forme $\hat{\theta}_n = \hat{\theta}(T_n)$. □

Remarque : l'EMV lui-même n'est pas forcément exhaustif. Considérons l'exemple suivant : $X \sim \mathcal{U}[\theta, 2\theta], \theta > 0$. La densité de X est donnée par

$$f(x, \theta) = \frac{1}{\theta} 1_{[\theta, 2\theta]}(x)$$

et la vraisemblance d'un $n-$échantillon de v.a. iid de même loi que X par

$$L_n(\theta) = \frac{1}{\theta^n} 1_{\theta \, \leq \, \inf_{1 \leq j \leq n} X_j \leq \sup_{1 \leq j \leq n} X_j \, \leq 2\theta}.$$

La statistique $(\inf_{1 \leq i \leq n} X_i, \sup_{1 \leq i \leq n} X_j)$ est exhaustive minimale pour θ. D'autre part, l'EMV $\hat{\theta}_n$ est par définition donné par la valeur de

$$\theta \in \Big[\sup_{1 \leq i \leq n} X_i/2, \inf_{1 \leq i \leq n} X_i \Big]$$

qui minimise θ^n (donc qui maximise $L_n(\theta)$). Remarquons de passage que $\sup_{1 \leq i \leq n} X_i/2 \leq \inf_{1 \leq i \leq n} X_i$ presque sûrement. On déduit que l'EMV est

$$\hat{\theta}_n = \frac{\sup_{1 \leq i \leq n} X_i}{2}$$

et que $\hat{\theta}_n$ ne peut être exhaustif pour θ.

On prouve aussi que l'EMV est invariant par reparamétrisation.

Théorème 5.1.2 *Pour n'importe quelle application g de Θ, $g(\hat{\theta}_n)$ est l'EMV de $g(\theta)$.*

Le théorème précédent est connu sous le nom de *Théorème de Zehna.*
Démonstration : On définit pour tout $\eta \in g(\Theta)$ la fonction

$$L_n^*(\eta) = \sup_{\theta : g(\theta)=\eta} L_n(\theta)$$

la vraisemblance de η. On a que

$$\begin{aligned} \sup_{\eta \in g(\Theta)} L *_n (\eta) &= \sup_{\theta \in \Theta} L_n(\theta) \\ &\quad \text{car on veut voir } \Theta \text{ comme l'union des ensembles } \{\theta \in \Theta \mid g(\theta) = \eta\} \\ &= L_n(\hat{\theta}_n). \end{aligned}$$

Or,

$$\begin{aligned} L_n(\hat{\theta}) &= sup_{\theta \ |g(\theta)=g(\hat{\theta}_n)} L_n(\theta) \\ &\quad \text{comme } \hat{\theta}_n \in \{\theta \mid g(\theta) = g(\hat{\theta}_n)\} \text{ et on sait qu'il réalise le sup de } L_n \\ &= L_n^*(g(\hat{\theta}_n)) \text{ par définition de } L_n^*. \end{aligned}$$

Il vient que $\sup_{\eta \in g(\Theta)} L_n^*(\eta) = L_n^*(g(\hat{\theta}_n))$ et donc $g(\hat{\theta}_n)$ (qui est clairement dans $g(\Theta)$) est bien l'EMV de $g(\theta)$. □

5.1.4 Propriétés asymptotiques de l'EMV

Soient $X_1, \ldots, X_n \sim f(x, \theta_0)$, $x \in \mathcal{X}$ et $\theta_0 \in \Theta$. On suppose que les hypothèses H0-H4 sont vérifiées.

Théorème 5.1.3 *Sous les hypothèse H0-H4 alors une suite $\hat{\theta}_n$ satisfaisant la condition du premier ordre existe P_{θ_0}-p.s. à partir d'un certain rang et converge vers θ_0 :*

$$\hat{\theta}_n \xrightarrow{p.s.} \theta_0 \quad \textit{lorsque } n \to +\infty.$$

En particulier, si l'EMV existe et est unique alors il est fortement convergeant.

Démonstration : Remarquons d'abord que si

$$\psi_n(\theta) = \log \frac{L_n(\theta)}{L_n(\theta_0)}$$

alors

$$\nabla_\theta \log L_n(\theta) = 0 \Leftrightarrow \nabla_\theta \psi_n(\theta) = 0.$$

D'après la loi forte des grands nombres

$$\frac{\psi_n(\theta)}{n} = \frac{1}{n}\sum_{j=1}^{n} \log \frac{f(X_i,\theta)}{f(X_i,\theta_0)} \xrightarrow{p.s.} \mathbb{E}_{\theta_0} \log\left[\frac{f(X,\theta)}{f(X,\theta_0)}\right] = -K(\theta_0,\theta).$$

Or $-K(\theta_0,\theta) \leq 0$ et $K(\theta,\theta_0)= 0$ si et seulement si $f(X,\theta) = f(X,\theta_0)$ avec probabilité 1. Or, d'après $H0$, c'est équivalent à $\theta = \theta_0$. Donc $\forall\ \theta \neq \theta_0$, il existe $N_\varepsilon \in \mathbb{N}$ tel que

$$P_{\theta_0}\left(\forall\ n > N_\varepsilon, \quad \psi_n(\theta) < 0\right) = 1.$$

Pour tout $\varepsilon > 0$ suffisamment petit pour que $[\theta_0 - \varepsilon; \theta_0 + \varepsilon] \subset \Theta$, il existe N_ε tel que

$$P_{\theta_0}(\forall\ n > N_\varepsilon, \quad \psi_n(\theta_0 \pm \varepsilon) < 0) = 1.$$

La fonction $\psi_n(\theta)$ étant continue, elle atteind son maximum sur $[\theta_0 - \varepsilon, \theta_0 + \varepsilon]$ compact. Soit $\hat{\theta}_n$ le point le plus proche de θ_0 pour lequel ce maximum est atteint. Par définition $\psi(\hat{\theta}_n) \geq \psi_n(\theta_0) = 0$ donc $\hat{\theta}_n$ ne peut être égal ni à $\theta_0 - \varepsilon$ ni à $\theta_0 + \varepsilon$ puisque $\psi_n(\theta_0 \pm \varepsilon) < 0$. Le maximum est réalisé en $\hat{\theta}_n$ à l'intérieur de l'intervalle donc $\hat{\theta}_n$ vérifie la condition du premier ordre sur ψ_n et donc aussi celle sur L_n. En résumé, $\forall \varepsilon > 0$ suffisamment petit, $\exists N_\varepsilon \in \mathbb{N}$ tel que

$$P_{\theta_0}\Big(\forall\ n > N_\varepsilon,\ \exists\ \hat{\theta}_n \text{ solution de l'équation de vraisemblance et } \ |\hat{\theta}_n - \theta_0| < \epsilon\Big) = 1.$$

En particulier, dès que $[\theta_0 - \varepsilon; \theta_0 + \varepsilon] \subset \Theta$ on a

$$P_{\theta_0}\Big(\forall\ n > N_\varepsilon,\ \exists\ \hat{\theta}_n \text{ solution de l'équation de vraisemblance}\Big) = 1,$$

donc à partir du rang N_ε la première assertion du théorème est prouvée. D'autre part, cette suite $\hat{\theta}_n$ vérifie pour tout $\varepsilon > 0$ $P_{\theta_0}(Q(\varepsilon)) = 1$, où $Q(\varepsilon) = \cap_{n>N_\varepsilon}\{\ |\hat{\theta}_n - \theta_0| < \varepsilon\}$. On vérifie que pour $\varepsilon < \delta$ alors $N_\delta \leq N_\varepsilon$ et $|\hat{\theta}_n - \theta_0| < \varepsilon$ implique $|\hat{\theta}_n - \theta_0| < \delta$. Donc $Q(\varepsilon) \subseteq Q(\delta)$ et pour tout $\epsilon_n \to 0$

$$P_{\theta_0}\left(\bigcap_{n\geq N_{\varepsilon_1}} \{|\hat{\theta}_n - \theta_0| < \varepsilon_n\}\right) = P_{\theta_0}\left(\bigcap_{n\in\mathbb{N}} Q(\varepsilon_n)\right) = \lim_{n\to\infty} P(Q(\varepsilon_n)) = 1.$$

Ceci implique que $P_{\theta_0}(\lim \hat{\theta}_n = \theta_0) = 1$ ce qui termine la preuve. □

Corollaire 5.1.1 *Si pour tout n l'estimateur il existe un unique $\hat{\theta}_n$ satisfaisant la condition du premier ordre alors la suite $(\hat{\theta}_n)_n$ est fortement convergente. De plus, si Θ est un intervalle $]\underline{\theta}, \overline{\theta}[$ pour $\underline{\theta}, \overline{\theta} \in (\mathbb{R} \cup \{\pm\infty\})^d$ alors cette solution coïncide P_{θ_0}-p.s. avec l'EMV à partir d'un certain rang.*

Démonstration : Le premier point est évident car alors $\hat{\theta}_n$ coïncide avec la suite construite dans la preuve du théorème 5.1.3 par unicité.
Supposons que Θ soit un intervalle et que $\hat{\theta}_n$, unique solution de la condition du premier ordre existe. Alors à partir d'un certain rang il réalise un maximum local d'après la preuve du théorème précédent. Montrons que c'est un maximum global. La fonction $\theta \mapsto \nabla_\theta l_n(\theta)$ s'annule en un unique point $\hat{\theta}_n$ de Θ. C'est une fonction continue donc elle est de signe constant de par et d'autre de $\hat{\theta}_n$ sur l'intervalle Θ. Autrement dit $\hat{\theta}_n$ est un extremum global. Mais c'est aussi un maximum local, donc c'est un maximal global et donc l'EMV. □

Remarque : La convergence presque sûre de l'EMV peut souvent se démontrer directement grâce à la LFGN. Par exemple, l'EMV de l'espérance d'une loi normale $\mathcal{N}(\mu, \sigma^2)$ est la moyenne emprique $\bar{X}_n$.

$$\bar{X}_n \xrightarrow{p.s.} \mu$$

d'après la loi forte des grands nombres mais aussi d'après le corollaire 5.1.1.

Le théorème suivant établit la loi asymptotique de n'importe quelle solution $\hat{\theta}_n$ de l'équation de vraisemblance qui vérifie $\hat{\theta}_n \xrightarrow{p.s.} \theta_0$. On aura cependant besoin des hypothèses supplémentaires suivantes :

H5 : Il existe $\varepsilon > 0$ tel que pour tout $1 \leq i, j \leq d$ une fonction $M_{i,j} \geq 0$ vérifie $\mathbb{E}_{\theta_0}[M_{i,j}(X)] < +\infty$ et

$$\left| \frac{\partial^2 \log f(x,\theta)}{\partial \theta_i \partial \theta_j} \right| \leq M_{i,j}(x), \quad \forall x \in \mathcal{X}, \quad \forall \theta \in \{\theta \in \Theta;\ \|\theta - \theta_0\| \leq \varepsilon\}.$$

H6 : la matrice $I(\theta_0)$ est définie positive.

Théorème 5.1.4 *Sous les hypothèses $H0 - H6$, tous $\hat{\theta}_n$ vérifiant la condition du premier ordre et $\hat{\theta}_n \xrightarrow{p.s.} \theta_0$ vérifient aussi*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(0, I^{-1}(\theta_0)).$$

En particulier, si l'EMV existe et est unique alors il est asymptotiquement efficace.

Démonstration : Soit $\hat{\theta}_n$ une suite de solutions de l'équation de vraisemblance. Posons

$$\varphi_n(\theta) = \frac{\sum_{j=1}^n S(X_j, \theta)}{n} = \frac{1}{n} \sum_{j=1}^n \nabla_\theta (\log f)(X_j, \theta).$$

Soit $1 \leq j \leq d$ un indice fixé. D'après le développement de Taylor à l'ordre 1 de φ_n au point $\hat{\theta}_n$, il existe $\bar{\theta}_n = (\bar{\theta}_{n,i})'_{1\leq i\leq d}$ vérifiant

$$0 = \varphi_{n,j}(\hat{\theta}_n) = \varphi_{n,j}(\theta_0) + \nabla(\varphi_{n,j})(\bar{\theta}_n)^T(\hat{\theta}_n - \theta_0) \text{ et } \bar{\theta}_{n,i} \in [\min(\theta_{0,i}, \hat{\theta}_{n,i}), \max(\theta_{0,i}, \hat{\theta}_{n,i})],$$

soit

$$\left(I_j(\theta_0) - I_j(\theta_0) - \nabla\varphi_{n,j}(\bar{\theta}_n)\right)^T(\hat{\theta}_n - \theta_0) = \varphi_{n,j}(\theta_0)$$

où I_j est le j-ème vecteur colonne de I. On sait que l'ensemble $\mathcal{C}(K)$ des fonctions continues munis sur $K = \{\theta \in \Theta;\ \|\theta - \theta_0\| \leq \varepsilon\}$ compact et muni de la norme uniforme $\|\cdot\|_K$ est un espace de Banach. Sous H5, on vérifie que

$$\mathbb{E}_{\theta_0}\|\partial^2 \log f(x,\theta)/\partial\theta_i\theta_j\|_K < \infty.$$

Si ∇_i est la dérivée par rapport à la i-ème coordonnée θ_i et $I_{i,j}$ et le coefficient i,j de l'information de Fisher, en appliquant la LFGN on obtient

$$P_{\theta_0}\left(\lim_{n\to\infty} \|\nabla_i\varphi_{n,j}(\theta) + I_{i,j}(\theta)\|_K = 0\right) = 1.$$

Mais la convergence uniforme sur K et la continuité de $\theta \to \nabla_i\varphi_{n,j}(\theta)$ entraîne la continuité de $\theta \to I_{i,j}(\theta)$ sur K. D'autre part, avec probabilité P_{θ_0} égale à 1, on sait que $\bar{\theta}_n$ converge vers θ_0 car $\hat{\theta}_n$ est fortement convergeant. En particulier, $\|\varphi_{n,j}(\bar{\theta}_n) + I_{i,j}(\theta_0)\| \leq \|\varphi_{n,j}(\bar{\theta}_n) + I_{i,j}(\bar{\theta}_n)\| + \varepsilon_n$ avec $\varepsilon_n \xrightarrow{p.s.} 0$ donc

$$\begin{aligned}\|\nabla_i\varphi_{n,j}(\bar{\theta}_n) + I_{i,j}(\theta_0)\| &\leq \|\varphi_{n,j}(\bar{\theta}_n) - I_{i,j}(\hat{\theta}_n)\| + \varepsilon_n \\ &\leq \|\nabla_i\varphi_{n,j}(\theta) - I_{i,j}(\theta)\|_K + \varepsilon_n \xrightarrow{p.s.} 0.\end{aligned}$$

Il s'en suit que pour tout $1 \leq j \leq d$ on a

$$(I_j(\theta_0) + o_{p.s.}(1))^T(\hat{\theta}_n - \theta_0) = \varphi_{n,j}(\theta_0)$$

où $o_{p.s.}(1)$ est un terme aléatoire qui tend vers 0 P_{θ_0}-p.s.. En écrivant la forme vectorielle de ce système d'équation et en multipliant par $\sqrt{n}$, on trouve

$$\sqrt{n}(I(\theta_0) + o_{\mathbb{P}}(1))(\hat{\theta}_n - \theta_0) = \sqrt{n}\varphi_n(\theta_0)$$

On conclut par le Théorème Central Limite appliqué aux vecteurs scores qui donne $\sqrt{n}\varphi_n(\theta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(0, I(\theta_0))$, le théorème de Slutsky et la δ-méthode sous H6. □

Exemple : Soient $X_1, \ldots, X_n$ iid d'une loi $Exp(\lambda)$.

$$L_n(\lambda) = \lambda^n e^{-\lambda\sum_{j=1}^n X_j}$$

$$l_n(\lambda) = n \log \lambda - \lambda \sum_{j=1}^{n} X_j.$$

L'unique estimateur satisfaisant la condition du premier ordre est

$$\hat{\lambda}_n = \frac{1}{\bar{X}_n}.$$

Les hypothèses H0-H6 étant vérifiées, on a d'après le théorème 5.1.4

$$\sqrt{n}(\hat{\lambda}_n - \lambda) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \lambda^2).$$

5.2 Méthode des moments

5.2.1 Définition

On se place dans le cadre d'observations unidimensionnelles, $X \in \mathbb{R}$. On suppose que pour un entier $r \geq 1$ donné $E[|X|^r] < +\infty$ (ce qui implique que $E[|X|^s] < +\infty \ \forall \ s \in [0, r]$). D'après la loi des grands nombres, on voit que quand $n \to +\infty$

$$M_n^s = \frac{1}{n} \sum_{j=1}^{n} X_j^s \xrightarrow{p.s.} E[X^s] = m_s$$

où $X_1, \ldots, X_n$ sont des variable aléatoires indépendantes de même loi que X. On rappelle que M_n^s s'appelle le moment empirique d'ordre s (cf. Chapitre 2).

Par conséquent, lorsque n est suffisamment grand le moment empirique M_n^s est proche du moment théorique $E[X^s]$. Une propriété similaire a lieu pour les moments centrés.

Cette propriété de convergence conduit à écrire l'égalité entre les moments théoriques et les moments empiriques de même ordre. Il faudrait toutefois que les moments théoriques choisis s'expriment effectivement en fonction des paramètres inconnus. On écrira autant d'égalités qu'il y a de paramètres inconnus.

Supposons que $\theta = (\theta_1, \ldots, \theta_d)' \in \mathbb{R}^d$. On considère ici les moments non-centrés. Pour obtenir un estimateur du vecteur θ, on résoudra d'abord le système d'équations

$$\begin{aligned} m_1 &= m_1(\theta) \\ m_2 &= m_2(\theta) \\ &\vdots \\ m_d &= m_d(\theta) \end{aligned}$$

où m_j est la valeur (inconnue) du moment théorique d'ordre j. La notation $m_j(\theta)$ signifie que l'espérance est bien calculée par rapport au modèle P_θ où θ est le vrai paramètre qu'on essaie d'estimer.

Si θ est la solution unique du système précédent, alors il existe une application bijective ϕ telle que $\theta = \phi(m_1, \ldots, m_d)$.

Définition 5.2.1 *On appelle estimateur de θ obtenu par la méthode des moments la solution*

$$\hat{\theta}_n = \phi(M_n^1, \ldots, M_n^d);$$

i.e., $\hat{\theta}_n$ est l'unique solution du système

$$\begin{aligned} M_n^1 &= m_1(\hat{\theta}_n) \\ M_n^2 &= m_2(\hat{\theta}_n) \\ &\vdots \\ M_n^d &= m_d(\hat{\theta}_n) \end{aligned}$$

Exemples :

- On observe un échantillon $(X_1, \ldots, X_n)$ de variables aléatoires i.i.d $\sim \mathcal{P}(\theta)$. On désire estimer le paramètre θ en utilisant la méthode des moments. Soit m_1 la valeur de l'espérance. On a

  $$m_1 = E(X).$$

  Or, $E(X) = \theta$. Par conséquent l'estimateur de θ par la méthode des moments est obtenu en remplaçant $E(X)$ par la moyenne empirique notée usuellement $\bar{X}_n$:

  $$\bar{X}_n = \hat{\theta}_n$$

  et donc $\hat{\theta}_n$ n'est autre que la moyenne empirique de l'échantillon.
- On observe un échantillon $(X_1, \ldots, X_n)$ de variables aléatoires iid $\sim \mathcal{N}(\theta, \sigma^2)$. Soient m_1 et m_2 les valeurs des moments d'ordre 1 et 2.

  $$\begin{aligned} m_1 &= E(X) = \theta \\ m_2 &= E(X^2) = \theta^2 + \sigma^2, \end{aligned}$$

  et donc en remplaçant m_1 et m_2 par la moyenne empirique et le moment empirique d'ordre 2, on obtient

  $$(\hat{\theta}_n, \hat{\sigma}_n) = \left(\bar{X}_n, \sqrt{M_n^2 - \bar{X}_n^2}\right)$$

  où $M_n^2 = 1/n \sum_{j=1}^n X_j^2$.

Comme il a été déjà indiqué dans l'introduction, un estimateur par la méthode des moments convergent presque sûrement vers le vrai paramètre de la distribution. Dans dans l'exemple précédent, cela garantit que l'inégalité $M_n^2 - \bar{X}_n^2 \geq 0$ soit satisfaite avec une probabilité égale à 1 lorsque $n \to \infty$. Dans le paragraphe suivant, nous énonçerons les propriétés asymptotiques des estimateurs par la méthode des moments.

5.2.2 Propriétés asymptotiques

Théorème 5.2.1 *Si θ est le vrai paramètre du modèle, et $\hat{\theta}_n = \phi(M_n^1, \ldots, M_n^d)$ est l'estimateur de θ par la méthode des moments tel que ϕ soit continue, alors*

$$\hat{\theta}_n \xrightarrow{p.s.} \theta, \quad n \to \infty.$$

L'estimateur par la méthode des moments est consistant.

Théorème 5.2.2 *Si les hypothèses du théorème 5.2.1 sont satisfaites et si en plus $E(|X|^{2d}) < +\infty$ et ϕ est différentiable, alors*

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma)$$

La matrice des covariances Σ est donnée par

$$\Sigma = (J_{(m_1,\ldots,m_d)}(\phi)) M (J_{(m_1,\ldots,m_d)}(\phi))^T$$

où

$$M_{ij} = m_{i+j}(\theta) - m_i(\theta) m_j(\theta) \quad 1 \leq i, j \leq d.$$

L'estimateur par la méthode des moments est asymptotiquement normal.

Remarque : L'estimateur par la méthode des moments n'est pas nécessairement asymptotiquement efficace.

5.3 Estimation Bayésienne

5.3.1 Deux visions différentes

Soit le modèle paramétrique $(\mathcal{X}, P_\theta), \theta \in \Theta$. Dans l'approche classique (que nous avons suivi jusqu'à maintenant), appelée aussi approche fréquentiste, l'espace des paramètres, Θ où on "cherche" le vrai paramètre du modèle P_θ, est un sous-ensemble de $\mathbb{R}^d$. L'approche fréquentiste repose sur la vision suivante : plus

le nombre d'observations $X_1, \ldots, X_n$ est grand, mieux on estime le paramètre inconnue θ déterminant la loi P_θ. Le principe de cette école repose d'une manière très générale sur la loi des grands nombres qui assure la convergence du rapport

$$\frac{\text{Nombre de fois qu'un évènement est observé}}{\text{Nombre d'essais}}$$

vers la vraie probabilité de cet événement. Pour illustrer ce concept, considérons l'expérience aléatoire du lancer d'un dé. La probabilité d'obtenir 3 (égale à 1/6 dans le cas d'un dé non truqué) est égale à la limite du ratio

$$R_n = \frac{\text{nombre de fois de 3 obtenus}}{n}, \quad n \to \infty$$

où n est le nombre de fois le dé a été lancé, c'est-à-dire, le nombre de fois que l'expérience aléatoire a été répétée. De fait, le point de vue fréquentiste est considéré par les non fréquentistes ou les Bayésiens comme limité puiqu'il ne donne pas la possibilité de calculer la probabilité d'un événement non répétable du genre "pleuvra t-il le 10 décembre 2010 ?" ou lorsque l'échantillon provient d'une population de taille finie (et alors $n \leq N$ ne peux tendre vers l'infini).

Dans l'approche Bayésienne, on supposera que le paramètre θ lui-même est la réalisation d'une variable aléatoire M définie sur un espace probabilisé $(\Lambda, \mathcal{L}, \pi_0)$ à valeurs dans l'espace mesurable $(\Theta, \mathcal{E})$. La loi π_0 est appelée la loi a priori. On supposera que l'espace $(\Theta, \mathcal{E})$ est muni d'une mesure positive σ-finie ν et que la loi π_0 admet une densité, souvent notée aussi π_0 et appelée densité à priori, relativement à ν.

Dans cette nouvelle approche, la loi de probabilité P_θ représente la loi conditionnelle de la variable aléatoire X sachant $M = \theta$. On supposera que l'espace probabilisé $(\mathcal{X}, \mathcal{B}, P_\theta)$ est dominé par une mesure σ-finie μ et que la loi P_θ admet une densité $f_{X|M=\theta}(x|\theta)$ relativement à μ ou tout simplement $f(x|\theta)$.

Dans ces conditions, le fait d'observer une réalisation x de la v.a. X apporte une information supplémentaire sur la loi de M. On s'intéressera donc à *la loi conditionnelle de M sachant que $X = x$*. Cette loi conditionnelle est appelée loi a posteriori de T. Soit $\pi(\theta|x)$ la densité de cette loi conditionnelle appelée densité à postériori.

La définition des densités conditionnelles nous permet d'écrire

$$f(x|\theta)\pi_0(\theta) = \pi(\theta|x)f(x), \quad x \in \mathcal{X}, \theta \in \Theta$$

où $f(x)$ est la loi marginale de X ; i.e.,

$$f(x) = \int_\Theta f(x|\theta)\pi_0(\theta)\nu(d\theta).$$

– Si la mesure dominante ν est égale à la mesure de Lebesgue, alors

$$f(x) = \int_\Theta f(x|\theta)\pi_0(\theta)d\theta.$$

– Si ν est égale à la mesure de comptage, alors

$$f(x) = \sum_{j:\theta_j\in\Theta} f(x|\theta_j)\pi_0(\theta_j)$$

où $\pi_0(\theta_j) = P(M = \theta_j)$.

5.3.2 Estimation Bayésienne

Supposons que l'on dispose de plusieurs observations $X_1, \ldots, X_n$ de la v.a. X. Dans le cadre Bayésien, les variables aléatoires X_j ne sont pas indépendantes. Cependant, on suppose qu'elles le sont conditionnellement à θ. Autrement dit, la loi conditionnelle de l'échantillon $(X_1, \ldots, X_n)$ admet pour densité (sachant $M = \theta$)

$$f(x_1, \ldots, x_n|\theta) = \prod_{j=1}^n f(x_j|\theta).$$

Proposition 5.3.1 (Formule de Bayes) *La densité de la loi conditionnelle de M sachant $(X_1, \ldots, X_n) = (x_1, \ldots, x_n)$ est donnée par*

$$\pi(\theta|x_1, \ldots, x_n) = \frac{\prod_{j=1}^n f(x_j|\theta)\pi_0(\theta)}{\int_\Theta \prod_{j=1}^n f(x_j|\theta)\pi_0(\theta)\nu(d\theta)}. \tag{5.3}$$

Démonstration : Pour tout $\theta \in \Theta$, la définition des densités conditionnelles nous permet d'écrire

$$f(x_1, x_2, \ldots, x_n|\theta)\pi_0(\theta) = h(\theta|x_1, \ldots, x_n)f(x_1, \ldots, x_n).$$

En intégrant des deux côtés sur Θ par rapport à la mesure dominante ν, on obtient

$$\begin{aligned} f(x_1, \ldots, x_n) &= \int_\Theta f(x_1, x_2, \ldots, x_n|\theta)\pi_0(\theta)\nu(d\theta) \\ &= \int_\Theta \prod_{j=1}^n f(x_j|\theta)\pi_0(\theta)\nu(d\theta). \end{aligned}$$

Il s'en suit que

$$\begin{aligned} \pi(\theta|x_1, \ldots, x_n) &= \frac{f(x_1, x_2, \ldots, x_n|\theta)\pi_0(\theta)}{\int_\Theta \prod_{j=1}^n f(x_j|\theta)\pi_0(\theta)\nu(d\theta)} \\ &= \frac{\prod_{j=1}^n f(x_j|\theta)\pi_0(\theta)}{\int_\Theta \prod_{j=1} f(x_j|\theta)\pi_0(\theta)\nu(d\theta)}. \end{aligned}$$

□

Remarque : La densité donnée par (5.3) est la densité a posteriori de M sachant $(X_1, \ldots, X_n) = (x_1, \ldots, x_n)$. Elle remplace la fonction de vraisemblance utilisée dans l'approche classique (fréquentiste).

• A. Estimateur de Bayes

Dans l'approche Bayésienne, la construction d'un estimateur sera basée sur la loi conditionnelle de M sachant $(X_1, \ldots, X_n) = (x_1, \ldots, x_n)$. On prendra souvent l'estimateur de Bayes.

Définition 5.3.1 *L'estimateur de Bayes est égale à l'espérance conditionnelle* $E[M|X_1 = x_1, \ldots, X_n = x_n]$ *donnée par*

$$\begin{aligned} E[M|X_1 = x_1, \ldots, X_n = x_n] &= \int_\Theta \theta \pi(\theta|x_1, \ldots, x_n)\nu(d\theta) \\ &= \frac{\int_\Theta \theta \prod_{j=1}^n f(x_j|\theta)\pi_0(\theta)\nu(d\theta)}{\int_\Theta \prod_{j=1}^n f(x_j|\theta)\pi_0(\theta)\nu(d\theta)}. \end{aligned}$$

Exemple. On suppose ici que l'espace des paramètres est $\Theta = \mathbb{R}$, que la v.a. M suit la loi uniforme $\mathcal{U}[0,1]$, que l'espace des observations est $\{0,1\}$ et que la loi conditionnelle de X sachant $M = \theta$ est $\mathcal{B}]\nabla\backslash(\theta)$. On prend μ égale à la mesure de comptage et ν à celle de Lebesgue. En utilisant la notation précédente, nous avons donc,

$$f(x|\theta) = \theta^x(1-\theta)^{1-x}, x \in \{0,1\} \quad \text{et} \quad \pi_0(\theta) = 1_{[0,1]}(\theta), \theta \in \mathbb{R}.$$

Considérons n observations $X_1, \ldots, X_n$ de la v.a. X. La loi conditionnelle de $(X_1, \ldots, X_n)$ sachant $M = \theta$ a pour densité

$$f(x_1, \ldots, x_n|\theta) = \prod_{j=1}^n \theta^{x_j}(1-\theta)^{1-x_j} = \theta^{\sum_{j=1}^n x_j}(1-\theta)^{n-\sum_{j=1}^n x_j}.$$

La loi marginale de $(X_1, \ldots, X_n)$ a pour densité

$$\begin{aligned} f(x_1, \ldots, x_n) &= \int_{\mathbb{R}} \theta^{\sum_{j=1}^n x_j}(1-\theta)^{n-\sum_{j=1}^n x_j}\ 1_{[0,1]}(\theta)d\theta \\ &= \int_0^1 \theta^{\sum_{j=1}^n x_j}(1-\theta)^{n-\sum_{j=1}^n x_j}\ d\theta \\ &= B\left(\sum_{j=1}^n x_j + 1, n - \sum_{j=1}^n x_j + 1\right) \end{aligned}$$

où B est la fonction Béta. Posons $\sum_{j=1}^{n} x_j = s_n$. L'expression précédente s'écrit encore

$$f(x_1, \ldots, x_n) = \frac{\Gamma(s_n+1)\Gamma(n-s_n+1)}{\Gamma(n+2)} = \frac{s_n!(n-s_n)!}{(n+1)!}.$$

Quant à la densité a posteriori de M, elle est donnée par

$$\begin{aligned} \pi(\theta|x_1, \ldots, x_n) &= \frac{\theta^{s_n}(1-\theta)^{n-s_n} 1_{[0,1]}(\theta)}{B(s_n+1, n-s_n+1)} \\ &= \begin{cases} \frac{\theta^{s_n}(1-\theta)^{n-s_n}}{B(s_n+1,n-s_n+1)} & \text{si } \theta \in [0,1] \\ 0 & \text{sinon.} \end{cases} \end{aligned}$$

L'estimateur de Bayes est donné donc par

$$\begin{aligned} E(M|X_1 = x_1, \ldots, X_n = x_n) &= \int_0^1 \frac{\theta^{s_n+1}(1-\theta)^{n-s_n} 1_{[0,1]}(\theta)}{B(s_n+1, n-s_n+1)}\, d\theta \\ &= \frac{B(s_n+2, n-s_n+1)}{B(s_n+1, n-s_n+1)} = \frac{s_n+1}{n+2}. \end{aligned}$$

Notons que cet estimateur est différent de l'estimateur classique s_n/n.

B. Estimateur de Bayes empirique

Dans la pratique, la loi de M dépend souvent d'un ou de plusieurs paramètres inconnus. La loi marginale de X dépendra donc aussi de ce(s) paramètre(s). Ceux-ci pourront être estimés dans le modèle paramétrique associé à la famille de ces lois marginales. Les estimateurs seront alors reportés dans l'expression de l'estimateur de Bayes.

Considérons l'exemple suivant. Soit X une variable aléatoire dont la loi conditionnelle sachant $M = \theta$ est une loi de Poisson de paramètre θ ; i.e.,

$$X|M = \theta \sim \mathcal{P}(\theta) \quad \text{ou encore} \quad f(x|\theta) = \frac{\theta^x}{x!} e^{-\theta}, x \in \mathbb{N}.$$

Nous supposons que la loi a priori de M est la loi exponentielle de paramètre $\lambda > 0$; i.e.,

$$\pi_0(\theta) = \lambda e^{-\lambda\theta} 1_{\theta \geq 0}.$$

Lorsqu'on observe un échantillon de taille n, sa densité marginale est donnée par

$$f(x_1, \ldots, x_n, \lambda) = \frac{\lambda}{\prod_{j=1}^{n} x_j!} \int_0^\infty \theta^{n\overline{x}_n} e^{-(n+\lambda)\theta} d\theta$$

où $\overline{x}_n = 1/n \sum_{j=1}^n x_j$. On reconnaît la forme d'une densité Gamma. Par conséquent, l'expression précédente peut s'écrire encore

$$f(x_1, \ldots, x_n, \lambda) = \frac{n\overline{x}_n!}{\prod_{j=1}^n x_j!} \ \frac{\lambda}{(n+\lambda)^{n\overline{x}_n+1}}.$$

La densité a posteriori de M est donc égale à

$$\pi(\theta, \lambda | x_1, \ldots, x_n) = \frac{(n+\lambda)^{n\overline{x}_n+1}}{n\overline{x}_n!} \ \theta^{n\overline{x}_n} e^{-(n+\lambda)\theta} \ 1_{\theta \geq 0}.$$

Par conséquent, la loi a posterior de M est une $\Gamma(n\overline{x}_n + 1, n + \lambda)$ et donc

$$E(M | X_1, \ldots, X_n) = \frac{n\overline{X}_n + 1}{n + \lambda}. \tag{5.4}$$

Pour estimer le paramètre inconnu λ, on peut utiliser la méthode de Maximum de Vraisemblance à partir du modèle marginal de densité $f(x_1, \ldots, x_n, \lambda)$:

$$l_n(\lambda) = \log f(X_1, \ldots, X_n, \lambda) = cte + \log \lambda - (n\overline{X}_n + 1) \log(n + \lambda)$$

$$\frac{dl_n(\hat{\lambda}_n)}{d\lambda} = \frac{1}{\hat{\lambda}_n} - \frac{n\overline{X}_n + 1}{n + \hat{\lambda}_n} = 0 \quad \Rightarrow \hat{\lambda}_n = \frac{1}{\bar{X}_n}.$$

On vérifie facilement que la condition du deuxième ordre est bien remplie, et que

$$\hat{\lambda}_n = \frac{1}{\bar{X}_n}$$

correspond bien à un maximum global.

Finalement, en remplaçant dans (5.4) λ par son estimateur $\hat{\lambda}_n$, on obtient l'estimateur de Bayes empirique

$$\frac{\bar{X}_n(\bar{X}_n n + 1)}{\bar{X}_n n + 1} = \bar{X}_n.$$

Chapitre 6

L'estimation par intervalle de confiance

6.1 Définition

Soit le modèle statistique $(\mathcal{X}, P_\theta), \theta \in \Theta$. On supposera ici que $\Theta \subset \mathbb{R}$. Cependant, les définitions et résultats qui suivront se généralisent sans problème au cas multidimensionnel. Nous allons introduire la notion d'estimation ensembliste pour un paramètre inconnu.

Soient deux statistiques $A_n = T_1(X_1, \ldots, X_n)$ et $B_n = T_2(X_1, \ldots, X_n)$ où $(X_1, \ldots, X_n)$ est un $n-$échantillon de X. On se fixe $\alpha \in [0, 1]$.

Définition 6.1.1 *On dira que* $[A_n, B_n]$ *est un intervalle de confiance de niveau* $1 - \alpha$ *pour* θ *si*

$$P_\theta(\theta \in [A_n, B_n]) = 1 - \alpha$$

La notation P_θ nous "rappelle" que la probabilité de l'événement

$$\begin{aligned}\{\omega \in \Omega : \theta \in [A_n, B_n]\} \quad \equiv \quad & \{\omega \in \Omega : T_1(X_1(\omega), \ldots, X_n(\omega)) \leq \theta\} \\ & \cap \{\omega \in \Omega : T_2(X_1(\omega), \ldots, X_n(\omega)) \geq \theta\}\end{aligned}$$

dépend de la loi de l'échantillon $(X_1, \ldots, X_n)$ qui dépend à son tour du paramètre (inconnu) θ.

Dans l'approche fréquentiste, on doit comprendre un intervalle de confiance de niveau $1 - \alpha$ comme un intervalle aléatoire qui a une probabilité $1 - \alpha$ de contenir le vrai paramètre θ et non comme un intervalle fixé auquel θ aléatoire appartient avec une probabilité $1 - \alpha$.

Remarques :

– Pour un seuil α donné, il n'existe pas une manière unique de définir un intervalle de confiance. En effet, si on écrit $\alpha = \alpha_1 + \alpha_2,\ \alpha_1, \alpha_2 \geq 0$ (il existe un nombre infini de telles décompositions), il suffirait de prendre A_n et B_n tels que $P_\theta(\theta \in]-\infty, A_n]) = \alpha_1$ et $P_\theta(\theta \in [B_n, +\infty[) = \alpha_2$ à condition bien sûr que les événements $\{A_n \leq \theta\}$ et $\{B_n \geq \theta\}$ soient disjoints. Souvent, on répartit les risques de manière symétrique, soit $\alpha_1 = \alpha_2 = \alpha/2$.
– Un intervalle de confiance de la forme $[A_n, B_n]$ s'appelle un intervalle de confiance bilatéral. Cependant, il peut arriver que la recherche d'un intervalle unilatéral s'avère plus pertinente. Il s'agira d'un intervalle de la forme $[C_n, \infty[$ ou $]-\infty, C_n]$ où $C_n = T(X_1, \ldots, X_n)$ est une statistique.

Exemple :

Soit $\alpha \in [0,1]$ et soient $X_1, \ldots, X_n$ n v.a. iid $\mathcal{N}(\mu, 1)$. On peut construire un intervalle de confiance $[A_n, B_n]$ de niveau $1-\alpha$ pour μ qui sera basé sur la moyenne empirique $\bar{X}_n$. Celle-ci suit une loi normale de moyenne μ et de variance $1/n$, et par conséquent

$$P_\mu\left(\sqrt{n}(\bar{X}_n - \mu) \in [-z_{1-\alpha/2}, z_{1-\alpha/2}]\right) = 1-\alpha$$

où $z_{1-\alpha/2}$ est le quantile de $\mathcal{N}(0,1)$ d'ordre $1-\alpha/2$. En inversant les inégalités, il vient que

$$P_\mu\left(\mu \in \left[\bar{X}_n - \frac{z_{1-\alpha/2}}{\sqrt{n}}, \bar{X}_n + \frac{z_{1-\alpha/2}}{\sqrt{n}}\right]\right) = 1-\alpha$$

et

$$[A_n, B_n] = \left[\bar{X}_n - \frac{z_{1-\alpha/2}}{\sqrt{n}}, \bar{X}_n + \frac{z_{1-\alpha/2}}{\sqrt{n}}\right]$$

est un intervalle de confiance (symétrique) de niveau $1-\alpha$ pour l'espérance μ.

Définition 6.1.2 *On dira que $[A_n, B_n]$ est un intervalle de confiance de niveau asymptotiquement égal à $1-\alpha$ pour θ si*

$$\lim_{n\to\infty} P_\theta(\theta \in [A_n, B_n]) = 1-\alpha$$

Exemple : Soient $X_1, \ldots, X_n$ iid $\sim Exp(\lambda)$. L'estimateur de Maximum de Vraisemblance de λ est $\bar{X}_n^{-1}$. Par le Théorème Central Limite, la δ-méthode et le théorème de Slutsky, on a

$$\sqrt{n}\bar{X}_n\left(\frac{1}{\bar{X}_n} - \lambda\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0,1).$$

Par conséquent,

$$[A_n, B_n] = \left[\frac{1}{\bar{X}_n} - \frac{z_{1-\alpha/2}}{\sqrt{n}\bar{X}_n}, \frac{1}{\bar{X}_n} + \frac{z_{1-\alpha/2}}{\sqrt{n}\bar{X}_n}\right]$$

est un intervalle de confiance de niveau asymptotiquement égal à $1-\alpha$ pour λ.

6.2 Fonctions pivotales

Définition 6.2.1 *Une fonction pivotale pour θ est une variable aléatoire $\phi(X_1, \ldots, X_n, \theta)$ dont la loi est indépendante de θ.*

Une fonction pivotale n'est rien d'autre qu'une fonction de l'échantillon $(X_1, \ldots, X_n)$ et du paramètre θ dont la loi ne dépend d'aucun paramètre.

Définition 6.2.2 *Une fonction asymptotiquement pivotale pour θ est une variable aléatoire $\phi(X_1, \ldots, X_n, \theta)$ qui converge en loi vers une variable aléatoire dont la loi est indépendante de θ.*

Pour $\alpha \in [0,1]$ donné, les fonctions pivotales sont très pratiques pour construire un intervalle de confiance de niveau $1-\alpha$. Il suffit de prendre les deux quantiles q_1 d'ordre α_1 et q_2 d'ordre $1-\alpha_2$ de la loi de $\phi(X_1, \ldots, X_n, \theta)$ et résoudre les inégalités :

$$\begin{aligned} q_1 &\leq \phi(X_1, \ldots, X_n, \theta) \\ \phi(X_1, \ldots, X_n, \theta) &\leq q_2. \end{aligned}$$

Exemples :

– Dans l'exemple précédent où $X_1, \ldots, X_n$ sont des v.a. iid $\sim \mathcal{N}(\mu, 1)$, la variable

$$\phi(X_1, \ldots, X_n, \mu) = \sqrt{n}(\bar{X}_n - \mu)$$

est une fonction pivotale pour μ de loi $\mathcal{N}(0,1)$. D'une manière générale, la fonction

$$\phi(X_1, \ldots, X_n, \mu) = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}$$

est une fonction pivotale pour μ de loi $\mathcal{N}(0,1)$ dans le cas où $X_1, \ldots, X_n$ sont iid $\sim \mathcal{N}(\mu, \sigma^2)$ et σ.

– Supposons maintenant que σ est inconnu. La variable aléatoire

$$\phi(X_1, \ldots, X_n, \mu) = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\left[(n-1)^{-1} \sum_{j=1}^{n} (X_j - \bar{X})^2\right]^{1/2}}$$

est une fonction pivotale pour μ suivant la loi de Student T_{n-1} à $n-1$ degrés de liberté.

– Soient $X_1, \ldots, X_n$ des v.a. iid $\sim \mathcal{B}(1, p)$. La variable aléatoire

$$\phi(X_1, \ldots, X_n, \mu) = \frac{\sqrt{n}(\bar{X}_n - p)}{\sqrt{p(1-p)}}$$

est une fonction asymptotiquement pivotale pour p de loi asymptotique $\mathcal{N}(0, 1)$.

Troisième partie

Tests

Chapitre 7

Tests paramétriques

7.1 Généralités

7.1.1 Quelques définitions

Soit X une variable aléatoire, fonction mesurable de $(\Omega, \mathcal{A}, P)$ dans $(\mathcal{X}, \mathcal{A}', P_\theta)$, $\theta \in \Theta$. Supposons que Θ est partitionné en deux sous-ensembles *non vides* Θ_0 et Θ_1 (*i.e.* $\Theta = \Theta_0 \cup \Theta_1$ tel que $\Theta_0 \cap \Theta_1 = \emptyset$).

L'objectif d'un test est de décider si $\theta \in \Theta_0$, ou pas. Les deux hypothèses appelées hypothèse nulle $H_0 : \theta \in \Theta_0$ et son hypothèse alternative $H_1 : \theta \in \Theta_1$ n'ont pas des rôles symétriques. On définit la zone de rejet ou région critique du test l'événement $R = \{X \in C\} \in \mathcal{A}$ où C est un élément de $\mathcal{B}$ à déterminer. On rejette l'hypothèse H_0 lorsque la réalisation de l'expérience vérifie $X(\omega) \in C$, et on l'accepte dans le cas contraire. Plus généralement,

Définition 7.1.1 *Un test est une fonction mesurable $\phi : X \to [0,1]$, on refuse l'hypothèse H_0 lorsque $\phi(X) = 1$ et on l'accepte lorsque $\phi(X) = 0$.*

Lorsque ϕ prend aussi des valeurs distinctes de 0 et de 1 on parlera de test randomisé et, lorsque $\phi(X) \in]0,1[$, on rejette l'hypothèse H_0 avec la probabilité $\phi(X)$.

Lorsque le test ϕ n'est pas randomisé, on appelle zone de rejet du test l'ensemble $R = \{\phi(X) = 1\}$ $(= (\phi \circ X)^{-1}(\{1\}))$.

Evidemment, une zone de rejet R permet de construire un test non randomisé donné par $\phi = \mathbb{1}_R$ qui vaut 1 ou 0 selon que $X \in C$ ou $X \notin C$.

Définition 7.1.2 (Cas des tests non-randomisés) *Le niveau du test aussi appelé risque de première espèce est la probabilité maximale de rejeter l'hypothèse H_0 à tort, $\alpha = \sup_{\theta \in \Theta_0} P_\theta(R)$.*

La puissance du test est la fonction $\beta : \Theta_1 \to [0,1]$ définie par $\beta(\theta) = P_\theta(R)$ lorsque $\theta \in \Theta_1$. Le test est sans biais si $\beta(\theta) \geq \alpha$ pour tout $\theta \in \Theta_1$.

Dans le cas d'un test randomisé, on définit alors $\alpha = \sup_{\theta\in\Theta_0} \mathbb{E}_\theta(\phi)$ et $\alpha \leq \mathbb{E}_\theta(\phi)$ pour tout $\theta \in \Theta_1$ lorsqu'il est sans biais. Sa puissance est donnée par $\mathbb{E}_\theta(\phi)$ pour tout $\theta \in \Theta_1$.

Pour tout test, on appelle risque de second espèce la valeur $\sup_{\theta\in\Theta_1}(1-\beta(\theta))$.

Définition 7.1.3 *Lorsque $\Theta_j = \{\theta_j\}$ on dit que l'hypothèse est simple. Dans le cas contraire, on dit que l'hypothèse est composite.*

7.1.2 Tests d'hypothèse simple contre hypothèse simple

Ici $\Theta_1 = \{\theta_1\}$ et $\Theta_2 = \{\theta_2\}$. Le modèle est alors dominé, par exemple par la mesure $\mu = P_{\theta_1} + P_{\theta_2}$. Si $f(x,\theta_0), f(x,\theta_1)$ désignent les 2 valeurs prises par la densité de P_X par rapport à μ, on appelle test du rapport de vraisemblance (abrégé en TRV) le test

$$\phi(x) = f\left(\frac{f(x,\theta_1)}{f(x,\theta_0)}\right)$$

pour toute fonction $f : \mathbb{R}^+ \to [0,1]$ croissante donnée.

Dans le cas où f est l'indicatrice d'un intervalle $[k,+\infty[$, le test est non randomisé, lorsque $f(t) = 0$ si $t < k$ et $f(t) = 1$ pour $t > k$, on obtient un test randomisé. On accepte ici l'hypothèse H_1 lorsque le rapport $f(x,\theta_1)/f(x,\theta_0)$ est grand, c'est-à-dire si θ_1 est plus vraisemblable que θ_0.

Exemples :

- Modèle gaussien iid $X_1,\dots,X_n \sim \mathcal{N}(\theta,1)$.
  Dans ce cas $f(x,\theta) = (2\pi)^{-n/2}\exp\left(-\frac{1}{2}\sum_{i=1}^n (x_i-\theta)^2\right)$, et on pose

  $$\log\frac{f(x,\theta_1)}{f(x,\theta_0)} = (\theta_1-\theta_0)\sum_{i=1}^n x_i - \frac{n}{2}(\theta_1^2-\theta_0^2)$$

  est une fonction croissante de $\overline{x} = \frac{1}{n}\sum_{i=1}^n x_i$. Si $\theta_1 > \theta_0$, on rejettera donc l'hypothèse $\theta = \theta_0$ lorsque la statistique $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i > k$.
- Pour le modèle de Bernoulli iid $X_1,\dots,X_n \sim b(\theta)$ (où le paramètre $\theta \in [0,1]$).
  Soit $x = (x_1,\dots,x_n) \in \{0,1\}^n$, si on pose $s = x_1 + \cdots + x_n$, dans ce cas $f(x,\theta) = \theta^s(1-\theta)^{n-s}$. Ainsi

  $$\frac{f(x,\theta_1)}{f(x,\theta_0)} = \left(\frac{1-\theta_1}{1-\theta_0}\right)^n \left(\frac{\theta_1}{\theta_0}\Big/\frac{1-\theta_1}{1-\theta_0}\right)^s$$

  est une fonction croissante de s lorsque $\theta_1 > \theta_0$, donc la région critique est de la forme $s \geq k$.
  Notons que $S \sim B(n,\theta)$ suit une loi binomiale. Le niveau de ce test s'écrit $\alpha = \sum_{k\leq j\leq n} C_n^j \theta_0^j(1-\theta_0)^{n-j}$ et sa puissance $\beta = \sum_{k\leq j\leq n} C_n^j\theta_1^j(1-\theta_1)^{n-j}$.

Cette expression prend un nombre fini de valeurs, le niveau du test ne peut être fixé de manière exacte dans ce cas. Pour $\theta_1 > \theta_0$, on déduit que le test est sans biais.
Pour parvenir à fixer le niveau α d'un test on considérera un test randomisé de la forme $\phi(s) = 0$ lorsque $s < k$, $\phi(s) = \gamma$ lorsque $s = k$ et $\phi(s) = 1$ lorsque $s > k$. Alors l'entier k est le plus petit entier tel que $(\tilde{\alpha} =) P_{\theta_0}(S > k) < \alpha$ et γ est choisi en sorte d'ajuster le niveau à α. Par définition, $\tilde{\alpha} + \gamma P_{\theta_0}(S = k) \geq \alpha$ et on pose $\gamma = (\alpha - \tilde{\alpha})/P_{\theta_0}(S = k)$.

7.1.3 Tests asymptotiques

On voit sur les exemples précédents qu'il n'est pas toujours facile de construire un test de seuil α donné. Une étude asymptotique conduira à souvent à des tests de mise en oeuvre très simple.

Définition 7.1.4 *Soit* $\left(\mathcal{X}^{(n)}, \mathcal{A}'^{(n)}, P_\theta^{(n)}\right)$ *une suite de modèles statistiques sur le même espace de paramètres* Θ *; l'observation correspondante est notée* $X^{(n)}$.

Le niveau asymptotique d'une suite de tests de Θ_0 *contre* Θ_1 *de région de rejet respective* R_n*, pour tout* $n = 1, 2, \ldots$ *est la limite (lorsqu'elle existe)*

$$\alpha = \lim_{n\to\infty} \sup_{\theta\in\Theta_0} P_\theta\left(X^{(n)} \in R_n\right)$$

Cette suite de tests est consistante si

$$\forall \theta \in \Theta_1 : \qquad \lim_{n\to\infty} P_\theta\left(X^{(n)} \in R_n\right) = 1$$

Le seul cas abordé par ce cours est celui d'expériences iid dont la loi est notée P_θ pour lesquelles on peut construire un échantillon sur tout espace produit $(E^n, \mathcal{E}^{\otimes n})$ pour tout $n \in \mathbb{N}$.

Étudions le cas des tests asymptotiques à hypothèse simple $H_0 : \theta = \theta_0$ contre $H_1 : \theta \neq \theta_0$ où $\theta = \mathbb{E}_\theta(X)$. Le principe repose sur le théorème de limite centrale. Soit $X_1, X_2, X_3, \ldots$ une suite iid de loi P_θ telle que et $\mathrm{Var}_\theta X_1 = 1$. Sa région critique s'écrit

$$|\overline{X} - \theta_0| \geq \frac{\varphi_{1-\alpha/2}}{\sqrt{n}}$$

où $P(|\mathcal{N}(0,1)| \geq \varphi_{1-\alpha/2}) = \alpha$.

Le niveau asymptotique de ce test (pour $\theta \in \Theta_0$) suit du théorème centrale limite :

$$P_{\theta_0}\left(\sqrt{n}|\overline{X} - \theta_0| \geq \sqrt{n}\frac{\varphi_{1-\alpha/2}}{\sqrt{n}}\right) = P_{\theta_0}\left(\sqrt{n}|\overline{X} - \theta_0| \geq \varphi_{1-\alpha/2}\right) \to_{n\to\infty} \alpha.$$

La puissance de ce test s'écrit pour $\theta \neq \theta_0$

$$P_\theta\left(\sqrt{n}|\overline{X} - \theta_0| \geq \varphi_{1-\alpha/2}\right) \to 1,$$

en effet la loi (faible) des grands nombres implique $\overline{X} - \theta_0 \to \theta - \theta_0$, en probabilité. Par contre la convergence précédente n'est pas uniforme en θ : pour le prouver notons que si $\theta_n - \theta_0 = o\left(\frac{1}{\sqrt{n}}\right)$ alors

$$P_{\theta_n}\left(\sqrt{n}|\overline{X} - \theta_0| \geq \varphi_{1-\alpha/2}\right) \to \alpha.$$

Lorsqu'on n'a plus $\mathrm{Var}_\theta X_1 = 1$, on remplacera les observations X_i par X_i/s_n où $s_n^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2$ désigne un estimateur consistant de la variance, pour conserver les mêmes propriétés asymptotiques du test.

Remarques :

- De la même manière on envisage un test de niveau asymptotique α pour l'hypothèse composite $\theta \leq \theta_0$; la région de rejet s'écrit alors

$$\left\{\sqrt{n}(\overline{X} - \theta_0) \geq \varphi_{1-\alpha}\right\}$$

- Dans le cas du test $H_0 : g(\theta) = \gamma_0$ sur la moyenne, on posera, pour un $\lambda > 0$ fixé,

$$\mathbb{R}_n = \{\theta \in \Theta \mid \|g(\theta) - \gamma_0\| \geq \lambda/\sqrt{n}\}.$$

7.2 Approche de Neyman-Pearson

7.2.1 Lemme de Neyman-Pearson

Cette section porte sur la comparaison de différentes procédures de tests pour les mêmes hypothèses H_0 et H_1.

Définition 7.2.1 *Soient ϕ_j $j = 1, 2$ deux tests de niveau $\leq \alpha$ pour tester l'hypothèse simple $H_0 : \theta \in \Theta_0$ contre $H_1 : \theta \notin \Theta_0$. Le test ϕ_1 est uniformément plus puissant (UPP) que le test ϕ_2 si $\beta_{\phi_1}(\theta) \geq \beta_{\phi_2}(\theta)$ pour tout $\theta \in \Theta_1$, où β_{ϕ_j} est la puissance du test $\phi_j, j = 1, 2$.*

Lorsque l'hypothèse alternative est simple on parlera simplement de test plus puissant. Dans le cadre de tests d'hypothèse simple contre hypothèse simple, le résultat suivant assure que le test du rapport de vraisemblance est optimal. Rappelons qui si

$$V(x) = \frac{f(x, \theta_1)}{f(x, \theta_0)},$$

un test randomisé du rapport de vraisemblance (TRV) s'écrit alors

- $\phi_{k,c}(x) = 1$ si $V(x) > k$,
- $\phi_{k,c}(x) = 0$ si $V(x) < k$ et
- $\phi_{k,c}(x) = c \in]0,1[$ si $V(x) = k$.

Lemme 7.2.1 (Neyman-Person)
a) Soit $\alpha > 0$. Si le TRV $\phi_{k,c}$ est un test de niveau α, alors il est UPP que tout autre test de niveau $\leq \alpha$.
b) Si $\alpha \in [0,1]$, il existe un TRV $\phi_{k,c}$ de niveau exactement α, il peut être randomisé.
c) Soit ϕ un test PP de niveau α alors, pour tout $\theta \in \Theta = \{\theta_0, \theta_1\}$, on a $P_\theta(\phi(X) \neq \phi_{k,c}(X), V(X) \neq k) = 0$.

Démonstration :
a) Ici $\mathbb{E}_{\theta_0}\phi_{k,c}(X) = \alpha$. Soit donc ϕ tel que $\mathbb{E}_{\theta_0}\phi(X) \leq \alpha$, on doit prouver que $\mathbb{E}_{\theta_1}(\phi_{k,c}(X) - \phi(X)) \geq 0$. Notons que

$$\begin{aligned}\Delta &= \mathbb{E}_{\theta_1}(\phi_{k,c}(X) - \phi(X)) - k\mathbb{E}_{\theta_0}(\phi_{k,c}(X) - \phi(X)) \\ &= \mathbb{E}_{\theta_0}(\phi_{k,c}(X) - \phi(X))(V(X) - k) + \mathbb{E}_{\theta_1}(\phi_{k,c}(X) - \phi(X))\mathbb{I}_{\{f(X,\theta_0)=0\}}\end{aligned}$$

Si $\phi_{k,c}(x) = 0$ alors $V(x) - k < 0$, et si $\phi_{k,c}(x) = 1$ alors $\phi_{k,c}(x) - \phi(x) \geq 0$ car $\phi(x) \in [0,1]$. Ainsi le premier terme de l'identité précédente est positif. Notons que $\alpha > 0$ implique $k < \infty$, par suite $\phi_{k,c}(x) = 1$ lorsque $f(x,\theta_0) = 0$ et le second terme de l'identité précédente est aussi positif.
Alors $\mathbb{E}_{\theta_1}(\phi_{k,c}(X) - \phi(X)) \geq k\mathbb{E}_{\theta_0}(\phi_{k,c}(X) - \phi(X)) \geq 0$.
b) Notons d'abord que les cas extrêmes sont couverts. Si $\alpha = 0$ lorsque $k = \infty$ donne le test PP, $\phi_{\infty,0}$. Si $\alpha = 1$ lorsque $k = 0$ donne le test PP, $\phi_{0,0}$. Si maintenant, $\alpha \in]0,1[$, $P_{\theta_0}(V(X) = \infty) = 0$ alors il existe $k < \infty$ tel que $P_{\theta_0}(V(X) > k) \leq \alpha$ et $P_{\theta_0}(V(X) \geq k) \geq \alpha$.
Lorsque $P_{\theta_0}(V(X) = k) = 0$, on peut choisir $c = 0$ et on obtient donc un test non randomisé.
Sinon, $c = (\alpha - P_{\theta_0}(V(X) > k))/P_{\theta_0}(V(X) = k) > 0$ donne lieu à un test PP est obtenu avec k défini plus haut.
c) se traite comme les points précédents. □

Pour conclure cette section, le lemme suivant nous donne une évaluation de la différence entre puissance et niveau d'un tel test, c'est à dire le biais de ce test.

Lemme 7.2.2 *Dans le modèle $\{\theta_0, \theta_1\}$ la mesure P_θ est dominée par P_{θ_0} et si α et β désignent le niveau et la puissance d'un test d'hypothèse simple contre hypothèse simple correspondant, alors $\beta - \alpha \leq \frac{1}{2}\int |f(x,\theta_0) - f(x,\theta_1)| d\mu(x)$.*

Démonstration : On écrit $\beta - \alpha = \int \phi(x)(f(x,\theta_1) - f(x,\theta_0))d\mu(x)$. Or cette expression est maximisée par le test $\phi(x) = \mathbb{1}_{f(x,\theta_0)<f(x,\theta_1)}$. On conclut avec la relation $\int_{f(x,\theta_1)>f(x,\theta_0)}(f(x,\theta_1) - f(x,\theta_0))d\mu = \frac{1}{2}\int |f(x,\theta_1) - f(x,\theta_0)|d\mu$. □

Exemples :

– On teste une hypothèse gaussienne simple, $N(\mu_0, \Sigma_0)$ contre $N(\mu_1, \Sigma_1)$ en rejetant H_0 lorsque $V(X)$ est grand. Les lois étant continues, on utilise des tests non randomisés. La zone de rejet s'écrit

$$Q = (X-\mu_0)^t\Sigma_0^{-1}(X-\mu_0) - (X-\mu_1)^t\Sigma_1^{-1}(X-\mu_1) > q \text{ (est grand)}$$

Lorsque $\Sigma_0 = \Sigma_1$ et $\mu_1 = \mu_0 + \lambda\Delta$ où $\|\Delta\| = 1$ et $\lambda \in \mathbb{R}$, on rejettera l'hypothèse H_0 si

$$\Delta^t\Sigma_0^{-1}(X-\mu_0) > \varphi_{1-\alpha}\Delta^t\Sigma_0^{-1}\Delta$$

La zone de rejet dépend ici de la direction Δ de la différence mais pas de l'amplitude λ. Par contre, la puissance de ce test en dépend largement.

– Si $N_1, \ldots, N_k$ désignent le nombre d'occurrences de $1, \ldots, k$ dans un n-échantillon de loi multinomiale $\mathcal{M}(k, \theta_1, \ldots, \theta_k)$, alors

$$f(n_1, \ldots, n_k, \theta) = \frac{n!}{n_1! \cdots n_k!}\theta_1^{n_1} \cdots \theta_k^{n_k}.$$

Ici $V(\theta^1, \theta^0) = \prod_{i=1}^k (\theta_1^1/\theta_i^0)^{N_i}$.
Pour tester une hypothèse simple

$$\theta^0 = (\theta_1^0, \ldots, \theta_k^0)/\theta^1 = (\theta_1^1, \ldots, \theta_k^1)$$

dans laquelle $\theta_1^0 > 0$ pour chaque i, on suppose l'alternative de la forme $\theta_i^1 = \epsilon\theta_0^i$ pour un $0 < \epsilon < 1$, et pour $i \neq j$ et $\theta_j^1 = \rho\theta_j^0 > 0$ avec $\rho = (1 - \epsilon\theta_j^0)/(1 - \theta_j^0)$.
Alors $V = \rho^n(\epsilon/\rho)^{N_j}$ et comme $\epsilon < 1$ implique $\rho \geq \epsilon$, on en déduit que la zone de rejet s'écrit $(N_j > k)$, ce qui signifie que l'on retourne au cas binomial déjà envisagé.

7.2.2 Rapports de vraisemblance monotones

Définition 7.2.2 *Soit $(P_\theta)_{\theta\in\Theta}$ un modèle μ-dominé avec $\Theta \subset \mathbb{R}$. On pose*

$$V_{\theta_1,\theta_2}(x) = \frac{f(x,\theta_2)}{f(x,\theta_1)}$$

Si $T(X)$ une statistique exhaustive le modèle est à rapport de vraisemblance monotone en T (RVM en T) lorsque $V_{\theta_1,\theta_2}(x)$, qui s'écrit comme fonction de $T(x)$, par exhaustivité, est une fonction croissante de $T(x)$ pour $\theta_1 < \theta_2$.

Exemple : Lorsque $\theta \mapsto g(\theta)$ est une fonction croissante le modèle exponentiel $f(x,\theta) = h(x)\exp\left(g(\theta)T(x) - B(\theta)\right)$ est à RVM en T.

Théorème 7.2.1 (Karlin-Rubin) *Soit $(P_\theta)_{\theta\in\Theta}$ un modèle à RVM en T, on considère le test randomisé $\phi_{t,c}(x) = 1$ lorsque $T(x) > t$, $\phi_{t,c}(x) = 0$ lorsque $T(x) < t$, et $\phi_{t,c}(x) = c$ si $T(x) = t$ pour $t \in \mathbb{R}$ et $c \in [0,1]$, fixés. Alors $\theta \mapsto \beta_\theta = \mathbb{E}_\theta\phi_{t,c}(X)$ est une fonction croissante. Si $\alpha = \mathbb{E}_{\theta_0}\phi_{t,c}(X) > 0$, le test $\phi_{t,c}$ est UPP parmi les tests sans biais de niveau α d'hypothèse composite $\theta \leq \theta_0$ contre $\theta > \theta_0$.*

Notons que celà implique que ce test UPP est aussi sans biais. Sa preuve est de même nature que celle du lemme 7.2.1. Donnons maintenant une variante de cet énoncé pour un test bilatère c'est-à-dire de la forme $\Theta_0 = \{\theta \in \Theta |\, \theta \leq \theta_1 \text{ ou } \theta \geq \theta_2\}$, ou $\Theta_0 = [\theta_1, \theta_2]$ pour $\theta_1 \leq \theta_2$.

Théorème 7.2.2 (Lehmann) *Si le modèle est exponentiel, admet une densité $f(x,\theta) = h(x)\exp\left(g(\theta)T(x) - B(\theta)\right)$ telle que l'application g soit strictement croissante sur $\Theta_0 =]-\infty, \theta_1] \cup [\theta_2, \infty[$, alors un test UPP parmi tous les tests sans biais de niveau α est défini par $\phi(x) = 1$ pour $T(x) \in]t_1, t_2[$, $\phi(x) = 0$ pour $T(x) \notin [t_1, t_2]$, et $\phi(x) = c_i$ pour $T(x) = t_i$ lorsque $i = 1, 2$. De plus les constantes t_i, c_i sont déterminées par les relations $\mathbb{E}_{\theta_i}\phi(X) = \alpha$ pour $i = 1, 2$.*

Il existe aussi un test UPP de même nature et de niveau α pour tester l'hypothèse $\theta \neq \theta_0$ contre $\theta = \theta_0$ (ou encore $\theta \in [\theta_1, \theta_2]$ contre $\theta \notin [\theta_1, \theta_2]$). Sa zone de rejet est de la forme $T(X) \notin]t_1, t_2[$ $(\phi(X) = 1)$ et $\phi(x) = c_i$ pour $T(X) = t_i$ $(i = 1, 2)$.

Exemple : Dans le cas d'un n-échantillon gaussien considéré plus haut, le test associé à la zone de rejet $\{|\overline{X} - \theta_0| \leq \varphi_{1-\alpha/2}/\sqrt{n}\}$ est UPP.

7.3 Tests du score et de Wald

7.3.1 Test du score

Ici nous considérons $\Theta \subset \mathbb{R}$.

Définition 7.3.1 *Soient ϕ_1 et ϕ_2 deux tests de niveau $\leq \alpha$ pour tester l'hypothèse $H_0 : \theta \leq \theta_0$ contre $H_1 : \theta > \theta_0$.*
Le test ϕ_1 est localement uniformément plus puissant (LUPP) que le test ϕ_2 si il existe un voisinage ouvert $V \ni \theta_0$ tel que $\beta_{\phi_1}(\theta) \geq \beta_{\phi_2}(\theta)$ pour tout $\theta \in]\theta_0, +\infty[\cap V$.
Le niveau local du test s'écrit $\alpha = \sup_{\theta\leq\theta_0, \theta\in V} \mathbb{E}_\theta\phi$.

Lemme 7.3.1 *Quand on ne considère que des tests réguliers, dans le sens où l'application $\theta \mapsto \mathbb{E}_\theta\phi$ est dérivable en θ_0 (intérieur à Θ), un test de niveau local α est LUPP si pour tout autre test ψ de même type : $\frac{\partial}{\partial\theta}\mathbb{E}_\theta\phi\big|_{\theta_0} \geq \frac{\partial}{\partial\theta}\mathbb{E}_\theta\psi\big|_{\theta_0}$.*

Démonstration : Par définition, pour tout $\theta > \theta_0$ et tout $\theta' \leq \theta_0$, $\theta, \theta' \in V$ on a

$$\mathbb{E}_\theta\phi - \mathbb{E}_\theta\psi = \mathbb{E}_{\theta'}\phi - \mathbb{E}_{\theta'}\psi + (\theta - \theta')\frac{\partial}{\partial\theta}\mathbb{E}_{\theta'}(\phi - \psi) + o(\theta - \theta') \text{ (si } \theta \to \theta').$$

En prenant le supremum en θ' pour le voisinage V suffisamment petit, on sort la différence des niveaux locaux des deux tests qui est nulle et on obtient $\mathbb{E}_\theta\phi \geq \mathbb{E}_\theta\psi$.□

Définition 7.3.2 *Le test fondé sur la statistique $S_X(\theta_0) = \frac{\partial}{\partial\theta}\log f(X, \theta_0)$ est appelé test de score. Il rejette l'hypothèse $\theta \leq \theta_0$ pour les grandes valeurs de $S_X(\theta_0)$.*

Théorème 7.3.1 *Tout test LUPP(α) régulier est un test du score qui vérifie $\phi(x) = 1$ lorsque $S_x(\theta_0) > kf(x, \theta_0)$. On peut aussi imposer que $\phi(x) = c$ soit constant sur l'ensemble où $S_X(\theta_0) = kf(x, \theta_0)$.*

Théorème 7.3.2 *Sous les hypothèses usuelles H1-H4 du chapitre 3, le test de région critique*

$$\left(S_{(X_1,\ldots,X_n)}(\theta_0) > \sqrt{nI_n(\theta_0)}\varphi_{1-\alpha}\right)$$

avec $S_{(X_1,\ldots,X_n)}(\theta_0) = \sum_{j=1}^1 S_{X_j}(\theta_0)$ est asymptotiquement LUPP, pour tout estimateur convergent $I_n(\theta_0)$ de l'information de Fischer $I(\theta_0)$.

Preuve. Ce résultat est fondé sur le théorème limite $S_{(X_1,\ldots,X_n)}(\theta_0)/\sqrt{nI(\theta_0)} \to \mathcal{N}(0, 1)$. □

Remarque : Dans le cas iid par additivité du score la région de confiance asymptotique s'écrit

$$\left\{\sum_{i=1}^n S_{X_i}(\theta_0) > \sqrt{\sum_{i=1}^n (S_{X_i}(\theta_0))^2}\varphi_{1-\alpha}\right\}.$$

7.4 Tests asymptotiques

Nous considérons maintenant un ensemble de paramètres $\Theta \subset \mathbb{R}^d$ (ouvert). Supposons que $\Theta_0 = \{\theta \in \Theta |\, g(\theta) = 0\}$ où la fonction $g : \Theta \to \mathbb{R}^k$ est différentiable et telle que le rang de $J_g(\theta)$ soit $k \leq d$ constant pour tout $\theta \in \Theta$. On dit que g est de plein rang.

La situation asymptotique considérée est celle d'observations iid $X^{(n)} = (X_1, \ldots, X_n)$ dans le modèle régulier $(P_\theta)_{\theta\in\Theta}$.

Définition 7.4.1

– *Soit $\widetilde{\theta}_n$ une suite d'estimateurs asymptotiquement efficace de θ,*

$$\sqrt{n}\left(\widetilde{\theta}_n - \theta\right) \to_{n\to\infty} \mathcal{N}_d\left(0, I^{-1}(\theta)\right), \qquad \text{sous la loi } P_\theta$$

Les tests de Wald fondés sur cette suite ont pour région de rejet

$$R_n : \xi_n^W > \chi^2_{k,1-\alpha}, \quad \text{avec } \xi_n^W = g(\widetilde{\theta}_n)^t \left(J_g(\widetilde{\theta}_n) I_n^{-1}(\widetilde{\theta}_n) \nabla g(\widetilde{\theta}_n)^t\right)^{-1} g(\widetilde{\theta}_n)$$

– *L'estimateur du maximum de vraisemblance contraint $\widehat{\theta}_n^c$ est l'EMV fondé sur $X^{(n)}$ sous l'hypothèse $H_0 : g(\theta) = 0$. Les tests du Score fondés sur cette suite ont pour région de rejet*

$$R_n : \xi_n^S > \chi^2_{k,1-\alpha}, \qquad \text{avec} \qquad \xi_n^S = \nabla L_n(\widehat{\theta}_n^c)^t I_n^{-1}(\widehat{\theta}_n^c) \nabla L_n(\widehat{\theta}_n^c)$$

L'exemple typique d'une suite d'estimateurs $\widetilde{\theta}_n$ est celui de l'EMV.

L'efficacité asymptotique de la suite d'estimateurs $(\widehat{\theta}_n)_n$, est à la base de ces tests, en effet

$$\sqrt{n}\left(\widehat{\theta}_n - \theta\right) \overset{P_{\theta_0}}{\to}_{n\to\infty} \mathcal{N}_d(0, I^{-1}(\theta_0))$$

implique,

$$n\left(\widehat{\theta}_n - \theta\right)^t I(\theta_0) \left(\widehat{\theta}_n - \theta\right) \overset{P_{\theta_0}}{\to}_{n\to\infty} \chi^2_d$$

Et, pour un estimateur consistant, $\widehat{\theta}_n$ de θ_0, le lemme de Slutsky implique bien

$$n\left(\widehat{\theta}_n - \theta\right)^t I(\widehat{\theta}_n) \left(\widehat{\theta}_n - \theta\right) \overset{P_{\theta_0}}{\to}_{n\to\infty} \chi^2_d.$$

De plus, le théorème des extrema liés s'écrit avec le Lagrangien $L_n(\theta) + g(\theta)^t\lambda$ donc $\nabla L_n(\widehat{\theta}_{0,n}) + \nabla g(\widehat{\theta}_{0,n})^t \widehat{\lambda}_n = 0$ conduit à

$$\xi_n^S = \frac{1}{n} \widehat{\lambda}_n^t \nabla g(\widehat{\theta}_{0,n}) I^{-1}(\widehat{\theta}_{0,n}) \nabla g(\widehat{\theta}_{0,n})^t \widehat{\lambda}_n$$

Ainsi $\xi_n^S - \xi_n^W \to 0$ en $P_\theta-$probabilité, si on prouve (cf. Monfort & Gouriéroux, 1996, page 556) :

$$\widehat{\lambda}_n/\sqrt{n} \sim -\left(\nabla g(\theta_0)^t I^{-1}(\theta_0) \nabla g(\theta_0)\right)^{-1} \sqrt{n} g(\widehat{\theta}_{0,n})$$

Proposition 7.4.1 *Sous les hypothèses usuelles, les suites de tests de Wald et du score sont de niveau asymptotique α et convergentes.*

Démonstration : Nous ébauchons le premier cas du test de Wald. Sous P_θ,

$$\sqrt{n}\left(g(\widetilde{\theta}_n) - g(\theta)\right) \to \mathcal{N}_k(0, A) \qquad \text{avec} \qquad A = \nabla g(\theta) I^{-1}(\theta) \nabla g(\theta)^t.$$

Par suite sous Θ_0, $g(\theta) = 0$ et on a $\sqrt{n} A^{-1/2} g(\widetilde{\theta}_n) \to \mathcal{N}_k(0, I_k)$. Ainsi $\xi_n^W = \|\sqrt{n} A^{-1/2} g(\widetilde{\theta}_n)\|^2 \to \chi^2_k$ sous Θ_0. Les résultats en découlent.

7.5 Tests fondés sur la vraisemblance

Dans le contexte d'un test $H_0 : \theta \in \Theta_0$ vs $H_1 : \theta \in \Theta_1$ pour un modèle dominé, nous posons

$$V(x) = \frac{\sup_{\theta\in\Theta_1} f(x,\theta)}{\sup_{\theta\in\Theta_0} f(x,\theta)}, \qquad \lambda(x) = \frac{\sup_{\theta\in\Theta} f(x,\theta)}{\sup_{\theta\in\Theta_0} f(x,\theta)} \tag{7.1}$$

Un test fondé sur la vraisemblance consiste à rejeter l'hypothèse Θ_0 pour les grandes valeurs de $V(X)$ ou, de manière équivalente, celles de $\lambda(X)$ (plus commode à calculer), quand on observe X. Ce test coïncide avec celui de Neyman-Pearson pour le cas d'une hypothèse simple ou dans le cas de rapports de vraisemblance monotones.

Posons $\widehat{\theta}_n$ et $\widehat{\theta}_n^c$, les estimateurs du maximum de vraisemblance non contraint et contraint de θ dans les modèles statistiques respectifs $(P_\theta)_{\theta\in\Theta}$, et $(P_\theta)_{\theta\in\Theta_0}$, alors

$$\lambda(x) = \frac{f(x,\widehat{\theta}_n)}{f(x,\widehat{\theta}_n^c)}, \qquad \log\lambda_n = \sum_{i=1}^n \log\lambda(x_i) = l_n(\widehat{\theta}_n) - l_n(\widehat{\theta}_n^c).$$

Définition 7.5.1 *Les tests du rapport de vraisemblance fondés sur la suite des n observations iid $(X_1,\ldots,X_n)$ ont pour région de rejet*

$$R_n : \xi_n^{RV} > \chi^2_{m,1-\alpha}, \quad \textit{avec} \quad \xi_n^{RV} = 2(l_n(\widehat{\theta}_n) - l_n(\widehat{\theta}_n^c)).$$

où $m = dim(\Theta) - dim(\Theta_0)$.

Proposition 7.5.1 *Sous les hypothèses usuelles, les suites de tests de rapport de vraisemblance sont de niveau asymptotique α et convergentes.*

7.5.1 Moyenne d'une gaussienne

On considère ici $X_1,\ldots,X_n$ iid de loi $\mathcal{N}(\mu,\sigma^2)$ et $\Theta = \mathbb{R}\times\mathbb{R}^{+*}$. Ici

$$f(x_1,\ldots,x_n,\theta) = \frac{1}{(\sigma\sqrt{2\pi})^n}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)^2\right)$$

Dans ce cas l'estimateur du maximum de vraisemblance de $\theta = (\mu,\sigma^2)$ sur Θ vaut $\widehat{\theta}_n = (\overline{x}_n, \widehat{\sigma}_n^2)$ avec $\overline{x}_n = \frac{1}{n}\sum_{i=1}^n x_i$ et $\widehat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^n (x_i-\overline{x})^2$ Lorsque l'on cherche à tester l'hypothèse $\Theta_0 = \{(\mu,\sigma^2)\in\Theta |\, \mu = \mu_0\}$, on a besoin de trouver $\widehat{\theta}_n^c$ l'estimateur du maximum de vraisemblance de θ sur Θ_0. Dans ce cas

$$\frac{\partial}{\partial\sigma^2}\sum_{i=1}^n \log f(x_i,\theta) = \frac{1}{2}\left(\frac{1}{\sigma^4}\sum_{i=1}^n (x_i-\mu_0)^2 - \frac{n}{\sigma^2}\right) = 0$$

et donc $\widehat{\theta}_n^c = (\mu_0, \widehat{\sigma}_0^2)$ avec $\widehat{\sigma}_0^2 = \widehat{\sigma}_n^2 + (\overline{x}_n - \mu_0)^2$, car

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_0)^2 - (\overline{x} - \mu_0)^2$$

Par suite, le fait que $\frac{1}{\widehat{\sigma}_0^2}\sum_i (x_i - \mu_0)^2 = \frac{1}{\widehat{\sigma}_n^2}\sum_i (x_i - \overline{x})^2$ implique immédiatement que $\log \lambda_n = \frac{n}{2}\log\left(\frac{\widehat{\sigma}_0^2}{\widehat{\sigma}_n^2}\right)$. Ainsi

$$\begin{aligned}\log \lambda_n &= \frac{n}{2}\log\left(1 + \left(\frac{\bar{x}_n - \mu_0}{\widehat{\sigma}_n}\right)^2\right) \\ &= \frac{n}{2}\log\left(1 + \frac{1}{n-1}\left(\frac{\sqrt{n}(\bar{x}_n - \mu_0)}{\widehat{s}_n}\right)^2\right)\end{aligned}$$

avec $\widehat{s}_n^2 = \frac{n}{n-1}\widehat{\sigma}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$. Alors, $log\lambda_n$ est une fonction croissante de $|T_n(x_1, ..., x_n)| = \sqrt{n}\left|\frac{\overline{x}-\mu_0}{\widehat{s}}\right|$, avec $T_n(X_1, \ldots, X_n) \sim t(n-1)$ la loi de Student de $n-1$ degrés de liberté.

On rejettera donc l'hypothèse $\mu = \mu_0$ au niveau α lorsque $|T_n| > t_{n-1,1-\alpha/2}$

Les tests unilatères sont obtenus de la même manière :

- pour tester $\mu \leq \mu_0$ contre $\mu > \mu_0$, on rejette l'hypothèse nulle au niveau α quand $T_n > t_{n-1,1-\alpha}$, et
- pour tester $\mu \geq \mu_0$ contre $\mu < \mu_0$, on rejette l'hypothèse nulle au niveau α quand $T_n < t_{n-1,\alpha}$.

7.5.2 Moyenne de deux échantillons gaussiens

A présent, on observe deux échantillons indépendants entre eux, et iid

$$X_1, \ldots, X_{n_1} \sim \mathcal{N}(\mu_X, \sigma^2) \qquad \text{et} \qquad Y_1, \ldots, Y_{n_2} \sim \mathcal{N}(\mu_Y, \sigma^2)$$

Ici $\Theta = \{\theta = (\mu_X, \mu_Y, \sigma^2) |\, \mu_X, \mu_Y \in \mathbb{R}, \sigma^2 > 0\} = \mathbb{R}^2 \times \mathbb{R}^{+*}$, ainsi

$$\log f(x, y, \theta) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\left(\sum_{i=1}^{n_1}(x_i - \mu_X)^2 + \sum_{i=1}^{n_2}(y_i - \mu_Y)^2\right)$$

Dans ce cas, l'estimateur du maximum de vraisemblance s'écrit $\widehat{\theta} = (\overline{x}_{n_1}, \overline{y}_{n_2}, \widehat{\sigma}_n^2)$ avec, à présent,

$$\overline{x}_{n_1} = \frac{1}{n_1}\sum_{i=1}^{n_1} x_i,\ \overline{y}_{n_2} = \frac{1}{n_2}\sum_{i=1}^{n_2} y_i,\ \widehat{\sigma}_n^2 = \frac{1}{n}\left(\sum_{i=1}^{n_1}(x_i - \overline{x})^2 + \sum_{i=1}^{n_2}(y_i - \overline{y})^2\right),$$

avec $n = n_1 + n_2$. Enfin, sur $\Theta_0 = \{\theta \in \Theta | \, \mu_X = \mu_Y = \mu, \ \mu \in \mathbb{R}\}$, l'estimateur du maximum de vraisemblance obtenu vaut $\widehat{\theta}_0 = (\widehat{\mu}_0, \widehat{\mu}_0, \widehat{\sigma}_0^2)$ où

$$\widehat{\mu}_0 = \frac{1}{n}\left(\sum_{i=1}^{n_1} x_i + \sum_{i=1}^{n_2} y_i\right), \quad \widehat{\sigma}_0^2 = \frac{1}{n}\left(\sum_{i=1}^{n_1}(x_i - \widehat{\mu}_0)^2 + \sum_{i=1}^{n_2}(y_i - \widehat{\mu}_0)^2\right)$$

Ainsi en ajoutant des identités découlant du développement de $(X_i - \widehat{\mu}_0)^2 = \big((X_i - \overline{X}) + (\overline{X} - \widehat{\mu}_0)\big)^2$, on obtient $\log \lambda_n = \frac{n}{2} \log \frac{\widehat{\sigma}_0^2}{\widehat{\sigma}_n^2}$. Le test fondé sur λ_n rejette donc l'hypothèse Θ_0 quand $|T_n|$ est grand avec

$$T_n = \sqrt{\frac{n_1 n_2}{n}} \, \frac{\overline{X}_{n_1} - \overline{Y}_{n_2}}{\widehat{S}_n} \sim t(n-2)$$

où

$$\widehat{S}_n^2 = \frac{n}{n-2}\widehat{\sigma}_n^2 = \frac{1}{n-2}\left(\sum_{i=1}^{n_1}(X_i - \overline{X})^2 + \sum_{i=1}^{n_2}(Y_i - \overline{Y})^2\right).$$

Pour montrer que cette variable a bien une loi de Student, on utilise le théorème de Cochran qui montre que les variables $\overline{X}_{n_1}/\sigma$, $\overline{Y}_{n_2}/\sigma$, $\sum_{i=1}^{n_1}(X_i - \overline{X})^2/\sigma^2$ et $\sum_{i=1}^{n_2}(Y_i - \overline{Y})^2/\sigma^2$ sont indépendantes et de lois respectives $\mathcal{N}(\mu_X/\sigma, 1/n_1)$, $\mathcal{N}(\mu_Y/\sigma, 1/n_2)$, $\chi^2_{n_1-1}$ et $\chi^2_{n_2-1}$. Donc, pour un test de niveau α, la region de rejet est donnée par $|T_n| > t_{n-2,1-\alpha/2}$.

Des tests de niveau α sont obtenus pour les hypothèses
- $\mu_X \leq \mu_Y$ avec la région de rejet $T_n > t_{n-2,1-\alpha}$,
- $\mu_X \geq \mu_Y$ avec la région de rejet $T_n < t_{n-2,\alpha}$.

On peut montrer que ces tests sont aussi ceux du rapport de vraisemblance.

7.5.3 Covariance de deux échantillons gaussiens

A présent la suite $(X_1, Y_1), \ldots, (X_n, Y_n)$ est iid selon la loi gaussienne bidimensionnelle $\mathcal{N}_2\left(\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}\right)$. Dans ce cadre gaussien (avec $\Theta \subset \mathbb{R}^5$), on voudrait tester l'indépendance des composantes X et Y c'est tester $H_0 : \rho = 0$ contre $H_1 : \rho \neq 0$. Ici

$$\begin{aligned}\sum_{i=1}^{n} \log p_\theta(x_i, y_i) &= -n \log\left(2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}\right) - \frac{1}{2((1-\rho^2)}\left(\frac{1}{\sigma_X^2}\sum_{i=1}^{n}(x_i - \mu_X)^2\right. \\ &\left. -\frac{2\rho}{\sigma_X\sigma_Y}\sum_{i=1}^{n}(x_i - \mu_X)(y_i - \mu_Y) + \frac{1}{\sigma_Y^2}\sum_{i=1}^{n}(y_i - \mu_Y)^2\right)\end{aligned}$$

Les équations du maximum de vraisemblance ont la solution $\widehat{\mu}_X = \overline{x}_n$, et $\widehat{\mu}_Y = \overline{y}_n$

$$\begin{aligned}\widehat{\sigma}_X^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2, \qquad \widehat{\sigma}_Y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y}_n)^2,\\ \widehat{\rho} &= \frac{1}{n\widehat{\sigma}_X\widehat{\sigma}_Y}\sum_{i=1}^{n}(x_i - \overline{x}_n)(y_i - \overline{y}_n)\end{aligned}$$

Sous l'hypothèse nulle $\Theta_0 = \{\theta \in \Theta |\, \rho = 0\}$, on trouve $\widehat{\theta}_0 = (\overline{x}_n, \overline{y}_n, \widehat{\sigma}_X^2, \widehat{\sigma}_Y^2, 0)$ et ainsi,

$$\log \lambda_n = -\frac{n}{2}\log(1 - \widehat{\rho}^2)$$

est une fonction croissante de $|\widehat{\rho}|$ mais aussi de $|T_n|$, avec $T_n = \sqrt{n-2}\widehat{\rho}/\sqrt{1-\widehat{\rho}^2}$.

Si $\rho = 0$, on peut montrer que $T_n \sim t(n-2)$ suit une loi de Student, ce qui permet de construire un test de niveau donné.

Chapitre 8

Tests non paramétriques

Dans cette section, on présente quelques tests dans un contexte non paramétrique.

8.1 Test du χ^2

8.1.1 Cas élémentaire

On considère une suite $\underline{X}_1, ..., \underline{X}_n$ de v.a. iid de loi multinomiale $\mathcal{M}(p_1, \ldots, p_k)$. Malheureusement pour le titre de la section, si $p = (p_1, \ldots, p_k)$ est fonction d'un paramètre θ, on est dans un modèle paramétrique $p(\theta)$.

On peut parler de cadre non paramétrique si k n'est pas connu, mais, traditionnellement le test présenté ci-dessous est classé parmi les tests non-paramétriques, une justification est fournie par l'exemple d'utilisation qui suit.

Théorème 8.1.1 *Supposons $p_1, \ldots, p_k > 0$. Soit $N_{j,n} = \sum_{i=1}^n X_{i,j}$ le nombre des occurences de 1 dans la j-ème classe pour $j = 1, ..., k$.*

$$\widehat{\chi}_n^2 = \sum_{j=1}^k \frac{(N_{j,n} - np_j)^2}{np_j} \xrightarrow{\mathcal{L}} \chi^2_{k-1}.$$

Démonstration : Posons $N_n = (N_{1,n}, \ldots, N_{k,n})$, alors le théorème de limite centrale vectoriel implique que $n^{-1/2}(N_n - np) \xrightarrow{\mathcal{L}} \mathcal{N}_k(0, \Sigma)$ où $\Sigma = \text{ diag}(p) - pp^t$ (*i.e.* $\Sigma_{i,j} = p_i - p_i^2$ si $i = j$ et $= -p_i p_j$ sinon).

Posons $\Delta = \text{ diag}(p)^{-1/2}$ la matrice diagonale d'éléments $p_1^{-1/2}, \ldots, p_k^{-1/2}$, alors $\Delta Z_n^{-1/2} \Delta (N_n - np) \to \mathcal{N}_k(0, \Delta^t \Sigma \Delta)$ où $\Delta^t \Sigma \Delta = I_k - \sqrt{p}\sqrt{p}^t$ est la matrice de projection orthogonale sur $\sqrt{p}^\perp$.

Considérons une matrice orthogonale (d'ordre k) telle que

$$U \begin{pmatrix} \sqrt{p_1} \\ \cdot \\ \cdot \\ \sqrt{p_k} \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ \cdot \\ 0 \end{pmatrix}$$

Alors $U\Delta\Sigma\Delta^t U^t = UU^t - (U\sqrt{p})(U\sqrt{p})^t$ s'écrit

$$U\Delta\Sigma\Delta^t U^t = \begin{pmatrix} 0 & 0 & 0 & \dots & \dots \\ 0 & 1 & 0 & \dots & \dots \\ 0 & 0 & 1 & 0 & \dots \\ 0 & \dots & 0 & 1 & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & \dots \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & I_{k-1} \end{pmatrix}$$

Par conséquent $\|U\Delta Z_n\|^2 = \|\Delta Z_n\|^2 \sim \chi^2_{k-1}$. □

On en déduit le <u>test du χ^2</u> $H_0 : p = p_0$ contre $H_1 : p \neq p_0$, qui rejette l'hypothèse H_0 lorsque

$$\widehat{\chi}^2_n = \sum_{j=1}^{k} \frac{(N_{j,n} - np_{j,0})^2}{np_{j,0}} \geq \chi^2_{k-1,1-\alpha}.$$

Ce test est de niveau asymptotique α. L'asymptotique est admise en pratique lorsque $n \min_j p_j \geq 5$ comme le confirment les remarques relatives aux variables binomiales dans le chapitre 1. Il est consistant car si $p \neq p_0$, lorsque $n \to \infty$, la loi des grands nombres implique l'équivalent presque sûr suivant

$$\begin{aligned} \sum_{j=1}^{k} \frac{(N_{j,n} - np_{j,0})^2}{np_{j,0}} &= n \sum_{j=1}^{k} \frac{\left(\frac{N_{j,n}}{n} - p_{j,0}\right)^2}{np_{j,0}} \\ &\sim n \sum_{j=1}^{k} \frac{(p_j - p_{j,0})^2}{p_{j,0}} \end{aligned}$$

et l'inégalité stricte $\chi^2(p, p_0) = \sum_{j=1}^{k} \frac{(p_j - p_{j,0})^2}{p_{j,0}} > 0$ implique que la statistique précédente équivalente à $n\chi^2(p, p_0)$ tend presque sûrement vers l'infini dans l'hypothèse alternative $p \neq p_0$ ce qui justifie la forme de la zone de rejet et prouve la consistance du test du χ^2.

Exemple :

On veut faire un test de l'hypothèse (non-paramétrique) globale, sur la loi marginale d'un échantillon iid réel : $P_U = P_0$ contre $P_U \neq P_0$. Si on décompose $\mathbb{R} = \bigcup_{j=1}^k A_j$ en une partition $A_1, \ldots, A_k$ alors la loi de X est différente de P_0 lorsque $P(U \in A_j) \neq P_0(A_j)$ pour un certain $j \in \{1, \ldots, k\}$. Alors le théorème 8.1.1 permet de tester cette hypothèse (grâce au test du χ^2) en posant $\underline{X_i} = (\mathbb{1}_{U_i \in A_1}, ..., \mathbb{1}_{U_i \in A_k})$. La question cruciale est alors le choix des classes. Une façon de faire en accord avec la règle $np \geq 5$ d'adéquation de la binomiale à une gaussienne est (lorsque les lois sont continues) de choisir k classes de même probabilité $p = 5/n$. Comme P_0 est donné, on divise donc $\mathbb{R}$ en k classes de même P_0-probabilité (aux problèmes de divisibilité près).

Une autre manière de procéder consisterait à diviser l'échantillon empirique (réordonné) en classes de même poids empirique.

Cet exemple permet, bien sûr de classer le test du χ^2 dans cette section non-paramétrique.

8.1.2 Test d'indépendance

Ici $X_i = (Y_i, Z_i)$ prend ses valeurs dans $\{y_1, \ldots, y_\ell\} \times \{z_1, \ldots, z_m\}$.

L'indépendance des variables Y et Z se traduit par la relation $p_{i,j} = q_i r_j$ si on pose $p_{i,j} = P(X_1 = (y_i, z_j))$, $q_i = P(Y_1 = y_i)$ et $r_j = P(Z_1 = z_j)$. Le paramètre $\theta = (q_1, \ldots, q_{\ell-1}, r_1, \ldots, r_{m-1}) \in \mathbb{R}^{\ell+m-2}$ du fait des restrictions naturelles des paramètres $\sum_i q_i = \sum_j r_j = 1$. Par suite le nombre de degrés de liberté vaut ici $D = \ell m - 1 - (\ell + m - 2) = (\ell - 1)(m - 1)$. La statistique précédente s'écrit avec des notations standard,

$$\begin{aligned}
\widehat{\chi}_n^2 &= n \sum_{i=1}^{\ell} \sum_{j=1}^{m} \frac{\left(\frac{N_{i,j}}{n} - \frac{N_{i\cdot}\, N_{\cdot j}}{n^2} \right)^2}{\frac{N_{i\cdot}\, N_{\cdot j}}{n^2}} \\
&= n \sum_{i=1}^{\ell} \sum_{j=1}^{m} \frac{\left(N_{i,j} - \frac{N_{i\cdot}\, N_{\cdot j}}{n} \right)^2}{N_{i\cdot}\, N_{\cdot j}}
\end{aligned}$$

Cette suite converge en loi vers une χ^2_D. Remarquons qu'ici l'estimateur du maximum de vraisemblance du vecteur $\theta \in \mathbb{R}^{\ell+m-2}$ s'écrit $\widehat{\theta} = (\widehat{q}_1, \ldots, \widehat{q}_{\ell-1}, \widehat{r}_1, \ldots, \widehat{r}_{m-1})$ avec $\widehat{q}_i = \frac{N_{i,\cdot}}{n}$ pour $1 \leq i < \ell$ et $\widehat{r}_j = \frac{N_{\cdot,j}}{n}$ pour $1 \leq j < m$.

8.2 Test de Kolmogorov Smirnov

On considère ici, $X_1, X_2, \ldots$, une suite iid à valeurs réelles et de fonction de répartition $F(x) = P(X_1 \leq x)$ et on pose

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{(X_i \leq x)}$$

sa fonction de répartition empirique (qui est comme toute fonction de répartition, croissante, continue à droite et admet une limite à gauche en tout point).

Théorème 8.2.1 (Glivenko-Cantelli) *On a, presque sûrement,*

$$\sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \to_{n \to \infty} 0$$

Remarque : Par simplicité, nous supposerons F strictement croissante et continue. Alors l'inverse F^{-1} de F a le sens commun, $F(X)$ a une loi uniforme sur $[0, 1]$ et $F^{-1}(U)$ suit la même loi que X_1 lorsque U est uniforme sur $[0, 1]$. Sans cette hypothèse, il reste exact que $F^{-1}(U)$ a la loi de X_1, par contre l'exemple de $X_1 \sim b(\frac{1}{2})$ prouve que $F(X_1)$ qui ne prend que trois valeurs, ne peut donc être uniforme.

Ce premier théorème 8.2.1 justifie l'idée de considérer la statistique $\|F_n - F_0\|_\infty$ pour tester une hypothèse du type $F = F_0$ contre $F \neq F_0$. Soit $\epsilon \downarrow 0$, le théorème 8.2.1 prouve que la suite de tests de cette hypothèse dont la zone de rejet s'écrit $\|F_n - F_0\|_\infty \geq \epsilon$ est consistante. Pour envisager le niveau d'un tel test, il faut connaître les quantiles approchés de la loi de $\|F_n - F_0\|_\infty$. Le résultat suivant prouve que cette loi ne dépend que de n, on pourra donc la tabuler après avoir simulé des variables uniformes.

Nous admettrons le (difficile) théorème limite suivant :

Théorème 8.2.2 *Soit $F = F_X$ la fonction de répartition de X, alors*

(Théorème de Kolmogorov)

où la fonction de répartition de $\mathcal{S}$ est

$$F_{\mathcal{S}}(x) = 1 - 2 \sum_{k=1}^{\infty} (-1)^{k-1} e^{-2k^2 x^2}, \quad x \geq 0.$$

Remarques :

– La variable aléatoire S a la même loi que le maximum d'un Pont Brownien B ; i.e. un processus Gaussien défini sur $[0,1]$ tel que

$$\begin{aligned} E[B(t)] &= 0, \forall\, t \in [0,1] \\ Cov(B(t), B(s)) &= E[B(t)B(s)] = \min(s,t) - st \quad \forall\, t, s \in [0,1]. \end{aligned}$$

– Le théorème de Kolmogorov intervient dans les problèmes de test d'adéquation essentiellement dans des méthodes non-paramétriques.

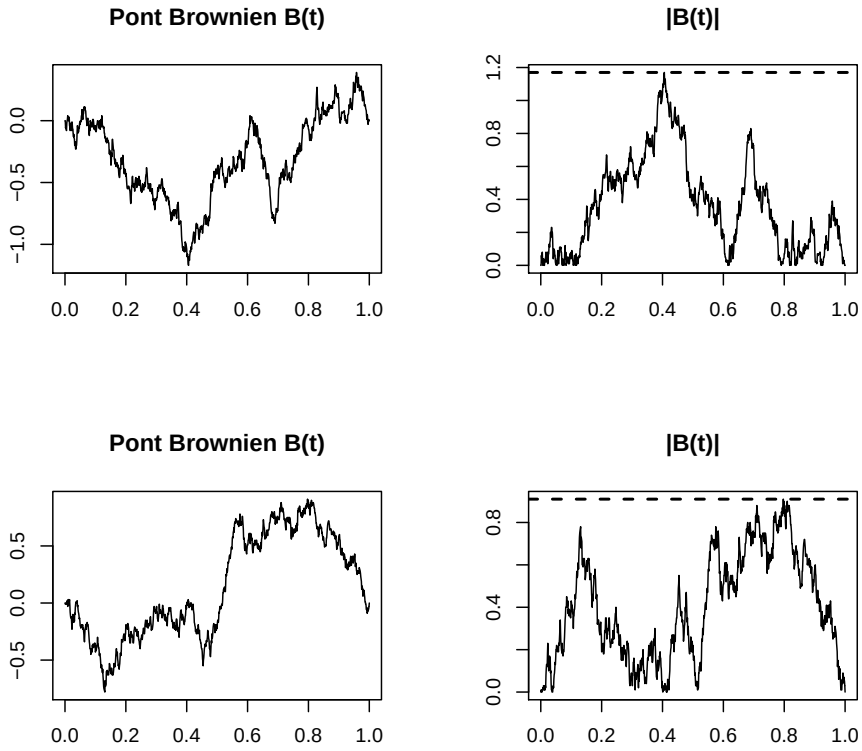

FIG. 8.1 – En première colonne : Les graphe de deux réalisations indépendantes d'un Pont Brownien $B(t)$. En deuxième colonne : les graphes de $|B(t)|, t \in [0,1]$. Les droites horizontales en pointillés passent par la valeur maximale $\max_{t\in[0,1]} |B(t)|$: 1,17 et 0,90 qui sont deux réalisations indépendantes de la variable aléatoire $\mathcal{S}$.

Théorème 8.2.3 (Kolmogorov & Smirnov) *Supposons que $F = F_0$. Les statistiques $D_n = \sqrt{n}\sup_x |F_n(x) - F_0(x)|$, $D_n^+ = \sqrt{n}\sup_x(F_n(x) - F_0(x))$, et $D_n^- = \sqrt{n}\sup_x(F_0(x) - F_n(x))$ ont une loi indépendante de F_0. De plus les lois de D_n^+ et D_n^- sont identiques.*

Les statistiques D_n, D_n^+ et D_n^- sont donc libre par rapport à F_0. Notons aussi que les variables aléatoires D_n^+ et D_n^- ont la même loi.

Théorème 8.2.4 (Smirnov et Kolmogorov) *On a respectivement*

$$\begin{aligned}\lim_{n\to\infty} P(D_n^+ > \lambda) &= \exp\left(-2\lambda^2\right) \qquad \text{et}\\ \lim_{n\to\infty} P(D_n > \lambda) &= 2\sum_{k=1}^{\infty}(-1)^{k+1}\exp\left(-2k^2\lambda^2\right)\end{aligned}$$

L'asymptotique est généralement admise dans le cas $n > 50$. Toutefois, il est raisonnable de vouloir comprendre le facteur $\sqrt{n}$. Ce lemme, très simple, est laissé en exercice au lecteur.

Lemme 8.2.1 *Posons $B_n(x) = \sqrt{n}(F_n(x) - F(x))$, alors pour tout $n-$uplet ordonné, $-\infty < x_1 \leq \cdots \leq x_k < \infty$, on a*

$$\begin{aligned}(B_n(x_1),\ldots,B_n(x_k)) &\to_{n\to\infty} (B_1,\ldots,B_k), \qquad \text{en loi}\\ (B_1,\ldots,B_k) &\sim \mathcal{N}_k(0,\Sigma),\\ \Sigma = (\sigma_{i,j})_{1\leq i,j\leq k}, \qquad \sigma_{i,j} &= F(x_i)\wedge F(x_j) - F(x_i)F(x_j)\end{aligned}$$

Il permet d'imaginer qu'un théorème de limite centrale "fonctionnel" gère le théorème 8.2.4, alors si on admet que $\sqrt{n}(F_n - F) \to B \circ F$ (en un sens non précisé, ici) pour un processus gaussien ([1]) centré B appelé "pont brownien", tel que $B(s) \sim \mathcal{N}(0, s - s^2)$, et tel que $\mathrm{Cov}(B(s), B(t)) = s \wedge t - st$ si $s, t \in [0, 1]$. Les lois du théorème 8.2.4 sont celles de $\sup_x B(x)$ et de $\|B\|_\infty$.

8.2.1 Test $F = F_0$

Pour tester les hypothèses $F = F_0$, $F \leq F_0$ ou $F \geq F_0$, on utilise les $(1-\alpha)$-quantiles $d_{n,1-\alpha}$ et $d^+_{n,1-\alpha}$ des lois de D_n, ou $D_n^\pm$ et on rejette l'hypothèse nulle lorsque la statistique adéquate dépasse le seuil correspondant.

– Pour tester $F = F_0$ contre $F \neq F_0$, on rejette l'hypothèse si $D_n > d_{n,1-\alpha}$,

[1] C'est à dire une famille de variables aléatoires, $B(t)$ pour $t \in \mathbb{R}$, telle que les combinaisons linéaires $\sum_{i=1}^I a_i B(t_i)$ aient toutes des lois gaussiennes ($\forall I$, $\forall a_i \in \mathbb{R}, \forall t_i \in [0,1]$, $i = 1,\ldots,I$).

– pour tester $F \leq F_0$ contre $F > F_0$, on rejette l'hypothèse si $D_n^+ > d_{n,1-\alpha}^+$,
– pour tester $F \geq F_0$ contre $F < F_0$, on rejette l'hypothèse si $D_n^- < d_{n,\alpha}^+$.

Les tests obtenus ont le niveau α et sont consistants.

Pour le montrer, on note, par exemple que lorsque

$$F < F_0 \Longrightarrow \limsup_n \sup_x (F_n(x) - F(x)) \leq 0$$

donc $P(\sup_x(F_n(x) - F(x)) > d) \to 1$ pour tout $d > 0$. Le comportement asymptotique de la suite $d_{n,1-\alpha}$ est obtenu en utilisant le théorème 8.2.4.

8.2.2 Cas de deux échantillons

On considère à présent deux échantillons réels indépendants iid $X_1, \ldots, X_n \sim F$ et $Y_1, \ldots, Y_m \sim G$. Les fonctions de répartition empiriques correspondantes sont notées F_n et G_m. Alors de manière analogue au tests de Kolmogorov Smirnov précédents, on peut démontrer

Théorème 8.2.5 *Posons* $c_{n,m} = \left(\frac{1}{n} + \frac{1}{m}\right)^{-1/2}$. *Les statistiques définies par les relations,* $D_{n,m} = c_{n,m} \sup_x |F_n(x) - G_m(x)|$, $D_{n,m}^+ = c_{n,m} \sup_x (F_n(x) - G_m(x))$, *et* $D_n^- = c_{n,m} \sup_x (G_m(x) - F_n(x))$ *ont des lois indépendantes de* F, G *si ces fonctions de répartitions sont continues et strictement croissantes.*

Cet énoncé permet aussi de simuler les quantiles de ces lois pour les tabuler.

Le but est de faire des tests pour les hypothèses

– $F = G$ contre $F \neq G$, la zone de rejet est $D_{n,m} > d_{n,m,1-\alpha}$,
– $F \leq G$ contre $F > G$, la zone de rejet est $D_{n,m}^+ > d_{n,m,1-\alpha}^+$, et
– $F \geq G$ contre $F < G$, la zone de rejet est $D_{n,m}^- < d_{n,m,\alpha}^-$.

Sous ces conditions, les suites $U_i = F(X_i)$ et $V_j = G(Y_j)$ sont iid et uniformes sur $[0, 1]$.

8.2.3 Ecriture en termes de statistique d'ordre

Les lois étant continues, la probabilité qu'il existe des ex-aequo dans cette liste est nulle.

Alors, on peut réécrire les expressions directement exploitables de ces statis-

tiques pour le cas de la comparaison des lois de deux échantillons,

$$\begin{aligned} D_{n,m} &= c_{n,m} \max \left\{ \left| \frac{i}{n} - \frac{j}{m} \right| \Big/ \ U_{(i)} < V_{(j)} < U_{(i+1)} \right\} \qquad (8.1) \\ &\text{et,} \quad \text{pour ses variantes signées,} \\ D^+_{n,m} &= c_{n,m} \max \left\{ \frac{i}{n} - \frac{j}{m} \Big/ \ U_{(i)} < V_{(j)} < U_{(i+1)} \right\}, \\ D^-_{n,m} &= c_{n,m} \max \left\{ \frac{j}{m} - \frac{i}{n} \Big/ \ U_{(i)} < V_{(j)} < U_{(i+1)} \right\} \end{aligned}$$

Et, pour les statistiques relevant d'un seul $n-$échantillon,

$$\begin{aligned} D_n &= \sqrt{n} \max \left\{ \left| \frac{i}{n} - F_0(u) \right| \Big/ \ U_{(i)} < F_0(u) < U_{(i+1)} \right\} \text{ et,} \qquad (8.2) \\ D^+_n &= \sqrt{n} \max \left\{ \frac{i}{n} - F_0(u) \Big/ \ U_{(i)} < u < U_{(i+1)} \right\}, \\ D^-_n &= \sqrt{n} \max \left\{ F_0(u) - \frac{i}{n} \Big/ \ U_{(i)} < u < U_{(i+1)} \right\} \end{aligned}$$

8.3 Tests de rang

8.3.1 Statistique de rangs

Définition 8.3.1 *Le rang de X_i dans la liste $X_1, \ldots, X_n$ vaut*

$$R_X(i) = 1 + \sum_{j \neq i} \mathbb{1}_{(X_j < X_i)}$$

C'est aussi le rang occupé par X_i lorsque cette liste est réordonnée de manière croissante, $X_{(1)} < X_{(2)} < \cdots < X_{(n)}$ où les $X_{(i)}$ sont les statistiques d'ordre.

Soit $(x_1, \ldots, x_n)$ un n-uplet de réels sans répétition alors l'application $i \mapsto R_x(i)$ ([2]), est injective $\{1, 2, \ldots, n\} \to \{1, 2, \ldots, n\}$; elle est donc bijective. Nous noterons encore cette bijection $R_x \in \mathcal{S}_n$. Rappelons que le groupe $\mathcal{S}_n$ des permutations de l'ensemble $\{1, 2, \ldots, n\}$ a le cardinal $n!$; sa structure algébrique est complexe ([3]).

De plus pour $x \in \mathbb{R}^n$ et $r \in \mathcal{S}_n$, on notera (avec précaution) $x_r = (x_{r_1}, \ldots, x_{r_n})$.

Plus globalement, l'application

$$\widetilde{\mathbb{R}^n} \to \mathcal{S}_n \times \mathbb{R}^n_<, \qquad (x_1, \ldots, x_n) \mapsto \big((R_x(1), \ldots, R_x(n)), (x_{(1)}, \ldots, x_{(n)}) \big)$$

[2] Elle associe à i, l'unique indice $j = R_x(i)$ de la statistique d'ordre vérifiant $x_{(i)} = x_j$.

[3] Ce groupe est non commutatif et il est simple pour $n > 4$.

est bijective sur l'ensemble $\widetilde{\mathbb{R}^n}$ des $n-$uplets distincts $(x_1, \ldots, x_n) \in \mathbb{R}^n$. Ici $\mathbb{R}^n_<$ désigne l'ensemble de $n-$uplets ordonnés $(u_1, \ldots, u_n) \in \mathbb{R}^n$ tels que $u_1 < \cdots < u_n$.

Cette situation est générique lorsque, comme nous le supposerons à partir de maintenant, la loi de $(X_1, \ldots, X_n)$ a une densité, $g(x_1, \ldots, x_n)$, par rapport à la mesure de Lebesgue sur $\mathbb{R}^n$. Alors les lois des statistiques de rang $R_X = (R_X(1), \ldots, R_X(n))$, et d'ordre $\Upsilon_X = (X_{(1)}, \ldots, X_{(n)})$ sont données par leur loi conditionnelle et leur densité

$$g_\Upsilon(\upsilon) = \sum_{r \in \mathcal{S}_n} g(\upsilon_r), \quad P(R_X = r |\ \Upsilon_X = \upsilon) = \frac{g(\upsilon)}{g_\Upsilon(\upsilon)}$$

Pour s'en convaincre, on note que l'événement $(\Upsilon_X \in B)$ s'écrit comme une partition, $(\Upsilon_X \in B) = \bigcup_{r \in \mathcal{S}_n} (\Upsilon_X \in B) \cap (R_X = r)$, avec

$$P\left((\Upsilon_X \in B) \cap (R_X = r)\right) = \int_B g(x_r) dx_r$$

Les tests fondés sur des statistiques de rang ont souvent pour hypothèse nulle celle que les variables $(X_1, \ldots, X_n)$ soient iid, lorsque la densité marginale vaut f, on a alors, $g(x_1, \ldots, x_n) = f(x_1) \cdots f(x_n)$ et le résultat suivant prouve l'intérêt de considérer les statistiques de rang.

Théorème 8.3.1 *Si le vecteur $(X_1, \ldots, X_n)$ est iid de densité marginale f par rapport à la mesure de Lebesgue, alors*

$$P(R_X = r) = \frac{1}{n!} g_\Upsilon(\upsilon)! \prod_{i=1}^n f(\upsilon_i)$$

Dans un modèle statistique non paramétrique indexé par f, R_X est une statistique libre et Υ_X est complète.

Remarque. Lorsque les lois ne sont plus continues, une manière de traiter les ex-aequo consiste à remplacer les rangs par les moyennes des rangs qu'ils occupent. Par exemple, dans la séquence $(1, \pi, 2, 5, \pi, 0)$ la suite des rangs pourrait s'écrire $(2, 4, 3, 6, 5, 1)$ ou $(2, 5, 3, 6, 4, 1)$, on lui préférera ici $(2, 4.5, 3, 6, 4.5, 1)$.

8.3.2 Statistiques linéaires de rang

Définition 8.3.2 *Soit $A = (a_{i,j})_{1 \le i,j \le n}$ une matrice réelle $n \times n$, la statistique linéaire de rang induite par la matrice A est*

$$L_A(X) = \sum_{i=1}^n a_{i, R_X(i)}$$

Théorème 8.3.2 *Si le vecteur X a des composantes iid,*

$$\mathbb{E}L_A(X)\overline{a}, \quad Var\, L_A(X) = \frac{1}{n-1}\sum_{i=1}^{n}\sum_{j=1}^{n}(a_{i,j} - a_{i,\cdot} - a_{\cdot,j} + \overline{a})^2$$

où

$$a_{i,\cdot} = \frac{1}{n}\sum_{j=1}^{n} a_{i,j}, \quad a_{\cdot,j} = \frac{1}{n}\sum_{i=1}^{n} a_{i,j}, \quad \overline{a} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} a_{i,j}$$

Preuve. Par l'équidisdribution des rangs,

$$\mathbb{E}L_A(X) = \sum_i \sum_h a_{i,h} P(R_X(i) = h)\overline{a}$$

Les définitions des coefficients liés à A impliquent en particulier que $\sum_i a_{\cdot,R_X(i)}\overline{a}$. Posons maintenant $\ell_i(h) = a_{i,h} - a_{\cdot,h} - a_{i,\cdot} - a_{\cdot,\cdot}$, il s'ensuit que $L_A(X) - \mathbb{E}L_A(X) = \sum_{i=1}^{n} \ell_i(R_X)$ et donc

$$\text{Var}\, L_A(X) = \sum_i \mathbb{E}L_i^2 + \sum_{i\neq j} \mathbb{E}L_i L_j, \qquad \text{avec} \qquad L_i = \ell_i(R_X(i))$$

On remarque d'abord que les expressions précédentes sont centrées $\mathbb{E}L_i = 0$. Utilisant l'équidistribution des rangs, le premier terme de cette somme, formé de termes diagonaux, est d'un calcul aisé,

$$\sum_i \mathbb{E}L_i^2 = \frac{1}{n}\sum_i \sum_h \ell_i^2(h)$$

A présent, la loi jointe de (L_i, L_j) s'obtient comme suit. La loi jointe des rangs $(R_X(i), R_X(j))$ s'écrit avec

$$P(R_X(i) = h, R_X(j) = k) = \begin{cases} \frac{1}{n(n-1)} & \text{lorsque } h \neq k \\ 0 & \text{si } h = k \end{cases}$$

Le couple $(R_X(i), R_X(j))$ ne peut en effet prendre que des valeurs distinctes et, une fois $R_X(i)$ choisi, il ne reste plus que $n-1$ valeurs envisageables pour $R_X(j)$.

Par suite,

$$\begin{aligned}\sum_{i\neq j}\mathbb{E}L_iL_j &= \frac{1}{n(n-1)}\sum_{i\neq j}\left(\sum_{h\neq k}\ell_i(h)\ell_j(k)\right)\\ &= -\frac{1}{n(n-1)}\sum_{i\neq j}\left(\sum_{h=1}^{n}\ell_i(h)\ell_j(h)\right)\\ &= -\frac{1}{n(n-1)}\sum_{h=1}^{n}\sum_{i=1}^{n}\ell_i(h)\left(\sum_{j\neq i}\ell_j(h)\right)\\ &= \frac{1}{n(n-1)}\sum_{h=1}^{n}\sum_{i=1}^{n}\ell_i^2(h)\end{aligned}$$

en vertu des relations, $\sum_{i=1}^n \ell_i(h) = 0$ et $\sum_{h=1}^n \ell_i(h) = 0$, déduites des définitions de $a_{i,\cdot}$, $a_{\cdot,j}$ et $\overline{a}$. Ainsi la relation $\frac{1}{n} + \frac{1}{n(n-1)} = \frac{1}{n-1}$ permet de conclure.

Remarques. Pour des statistiques linéaires simples, les expressions précédentes s'écrivent un peu mieux. Soient $A = a\alpha^t = (a_i\alpha_j)_{1\leq i,j\leq n}$ et $B = b\beta^t = (b_i\beta_j)_{1\leq i,j\leq n}$ on obtient, en posant $\overline{a} = \frac{1}{n}\sum_{i=1}^n a_i$ *(etc...)*

$$\mathbb{E}L_A(X)\overline{a\alpha}, \quad \operatorname{Var} L_A(X) = \frac{1}{n-1}\sum_i (a_i-\overline{a})^2 \sum_j (\alpha_j - \overline{\alpha})^2$$

Par bilinéarité de la variance, nous obtenons enfin

$$\operatorname{Cov}\left(L_A(X), L_B(X)\right) = \frac{1}{n-1}\sum_i (a_i-\overline{a})(b_i-\overline{b})\sum_j (\alpha_j-\overline{\alpha})(\beta_j-\overline{\beta})$$

Exercice. Indiquez comment tester l'hypothèse $m \leq m_0$ contre $m > m_0$ dans un modèle iid (on prouvera que l'on peut se ramener à un test de type pile ou face).

8.3.3 Test de Wilcoxon

Encore une fois, nous supposons que deux échantillons réels indépendants iid $X_1, \ldots, X_n \sim F$ et $Y_1, \ldots, Y_m \sim G$ ont des fonctions de répartition continues et strictement croissantes F et G.

L'objectif est de tester si $F = G$.

On pose $N + m$ et $(Z_1, \ldots, Z_N) = (X_1, \ldots, X_n, Y_1, \ldots, Y_m)$. On considère les rangs et statistiques d'ordre attachés à ces échantillons concaténés,

$$Z_{(1)} < Z_{(2)} < \cdots < Z_{(N-1)} < Z_{(N)}, \quad R_Z(i) = 1 + \sum_{j\neq i} 1_{(Z_j<Z_i)}, \quad 1 \leq i \leq N$$

Alors, R_Z est la permutation de $\{1,\dots,N\}$ telle que $Z_{R_Z(i)} = Z_{(i)}$. Cette variable aléatoire a une loi uniforme sur l'ensemble $\mathcal{S}_n$ des permutations de $\{1,\dots,N\}$ (de cardinal $N!$).

Définition 8.3.3 *La somme des rangs des X_i, $W_n = \sum_{i=1}^n R_Z(i)$ est appelée statistique de Wilcoxon.*

La loi de W_n (qui dépend de n et m) est tabulée. Notons que l'on peut toujours échanger les rôles de n et m à condition de remplacer W_n par une somme de $n+1$ à N, donc les tables ne comportent que le cas $n \leq m$. Evidemment, cette loi ne dépend pas de la loi F si $F = G$.
Un test pour l'hypothèse $F = G$ contre $F > G$ est donné par la zone de rejet $W_n > w_\alpha$. Ici w_α est le $(1-\alpha)-$quantile de la loi de W qui peut être tabulé en considérant car cette variable a la même loi (sous l'hypothèse nulle) que $W_U = \sum_{i=1}^n R_U(i)$ pour un échantillon aléatoire iid $U = (U_1,\dots,U_N)$ de marginales uniformes sur $[0,1]$ (*i.e.* $P(W_U > w_\alpha) = \alpha$).
Lorsque $n = 1$, la loi de W_1 est une loi de Bernoulli de paramètre $p = P(X_1 < Y_1)$; si $F = G$ alors $p = \frac{1}{2}$ et si $F > G$ alors $p = \int F(x)g(x)dx > \int G(x)g(x)dx = \frac{1}{2}$ ce qui permet de justifier la forme de la zone de rejet.

On a aussi
- $\mathbb{E}W_n \mathbb{E}R_Z(1) \sum_j \frac{j}{N} = \frac{n(N+1)}{2}$ (car $P(R_Z(i) = j) = \frac{1}{N}$)
- $\operatorname{Var} W_n = \frac{n(N+1)(N-n)}{12}$ (cf. théorème 8.3.2).

Ceci justifie (un peu) l'énoncé $\dfrac{W_n - \mathbb{E}W_n}{\sqrt{\operatorname{Var} W_n}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$ que nous ne prouverons pas ici.

8.3.4 Test de Spearman

Maintenant, $(X_1,Y_1),\dots,(X_1,Y_n)$ est une suite iid et on désire tester l'indépendance des X et des Y. On utilise la statistique de Spearman

$$S_n = \sum_{i=1}^n R_X(i) R_Y(i)$$

Sous l'hypothèse nulle, on obtient

$$\mathbb{E}S_n = \frac{1}{4}n(n+1)^2, \qquad \operatorname{Var} S_n = \frac{1}{144}(n-1)n^2(n+1)^2$$

Notons que les deux situations extrêmes, $R_X = R_Y$ et $R_X + 1 - R_Y$, conduisent à l'encadrement

$$\sum_i i(n+1-i)\frac{1}{6}n(n+1)(n+2) \leq S_n \leq \sum_i i^2 = \frac{1}{6}n(n+1)(2n+1)$$

Lorsque $n \to \infty$, cette distribution est asymptotiquement gaussienne ; donc une région critique du test de Spearman a la forme $(S_n < \underline{s}) \cup (S_n > \overline{s})$, pour un s tabulé permettant d'atteindre tout niveau α.

Enfin, la corrélation empirique des vecteurs aléatoires R_X et R_Y s'écrit

$$\rho_S = \frac{\frac{1}{n}\sum_i R_X(i)R_Y(i) - \frac{1}{n^2}\sum_i R_X(i)\sum_i R_X(i)}{V_X V_Y}$$

avec

$$\begin{aligned} V_X^2 &= \frac{1}{n}\sum_i R_X^2(i) - \left(\frac{1}{n}\sum_i R_X(i)\right)^2 \\ &= \frac{1}{n}\sum_i i^2 - \left(\frac{1}{n}\sum_i i\right)^2 = V_Y^2 \\ &= \frac{(n+1)(2n+1)}{6} - \left(\frac{n+1}{2}\right)^2 = \frac{n^2-1}{12} \end{aligned}$$

Par suite

$$\rho_{S_n} = \frac{12S_n - 3n(n+1)^2}{n(n^2-1)}$$

est une fonction affine du coefficient de Spearman, ce qui justifie d'introduire S_n pour tester une indépendance.