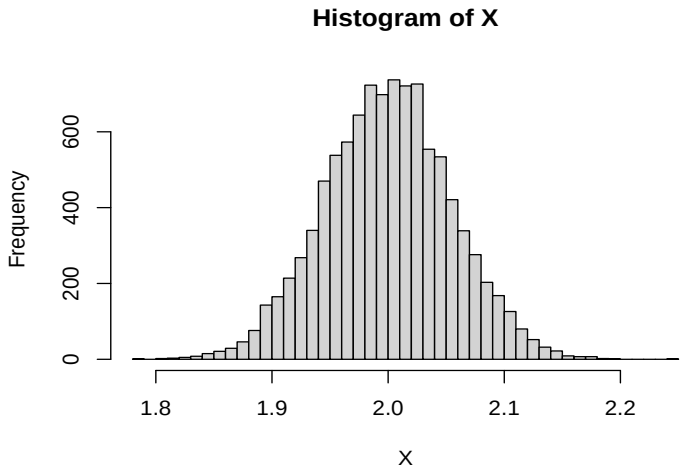
Université Paris I, Panthéon - Sorbonne

Licence M.I.A.S.H.S. Troisième Année

Cours de Statistique 2

Jean-Marc Bardet (Université Paris 1, SAMM)



U.F.R. 27 et SAMM (Statistique, Analyse et Modélisation Multidisciplinaire)
Université Panthéon-Sorbonne, 90 rue de Tolbiac, 75013 Paris.

Plan du cours

Introduction

1. Variables aléatoires et espérance
2. Vecteurs aléatoires et indépendance
3. Vecteurs gaussiens
4. Convergence et théorèmes limite
5. Estimation paramétrique
6. Tests paramétriques et non paramétriques

References

[1] Barbe et Ledoux, *Probabilités*, Belin.

[2] Dacunha-Castelle et Duflo, *Probabilités et Statistiques (I)*, Masson

[3] Dauxois, J. et Hassenforder, C. (2004). Toutes les probabilités et Statistiques. Cours et Exercices corrigés. Ellipses.

[4] Garet, O. et Kuntzmann, A., *De l'intégration aux probabilités*, Ellipses.

[5] Leboeuf, C., Guegand, J., Roque, J.L. et Landry, P. Cours de Probabilités et de statistiques, Ellipses

[6] Leboeuf, C.,Guegand, J.,Roque, J.L. et Landry, P. Exercices corrigés de probabilités, Ellipses

[7] Ross, S.M (2007). Initiation aux probabilités, Enseignement des Mathématiques. Presses polytechniques et universitaires romandes.

[8] Saporta, G. Probabilités, analyse des données et statistique (2nd édition), éditions Technip.

Introduction

Il demeure des choses inconnues à partir des connaissances antérieures en probabilités :

- Qu'est-ce qu'un événement et l'ensemble de tous les événements ?
- Que se passe-t-il pour des probabilités d'événements moins classiques (par exemple l'ensemble des décimaux) ?
- Comment traiter une variable aléatoire qui est continue et discrète à la fois (par exemple le nombre de minutes passées devant la TV) ?

Rappels: Mesures

Tribus

Notation.
- Ω est un ensemble (fini ou infini).
- $\mathcal{P}(\Omega)$ est l'ensemble de tous les sous-ensembles (parties) de Ω.

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

Définition. 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:

- $\Omega \subset \mathcal{F}$;
- lorsque $A \in \mathcal{F}$ alors $(\Omega \setminus A) \in \mathcal{F}$;
- pour tout $n \in \mathbb{N}^*$, lorsque $(A_1, \cdots, A_n) \in \mathcal{F}^n$ alors $A_1 \cup \cdots \cup A_n \in \mathcal{F}$.

Définition. 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 :

- $\Omega \subset \mathcal{F}$;
- lorsque $A \in \mathcal{F}$ alors $(\Omega \setminus A) \in \mathcal{F}$;
- pour $I \subset \mathbb{N}$, lorsque $(A_i)_{i \in I} \in \mathcal{F}^I$ alors $\bigcup_{i \in I} A_i \in \mathcal{A}$.

Exemple.

- Cas du Pile ou Face.
- Cas où Ω est infini : $\Omega = \mathbb{N}$ par exemple.

Propriété. *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 $I \subset \mathbb{N}$, 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. 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).

Remarque.

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

Rappel.

- Un ensemble ouvert U dans un espace métrique X est telle que pour tout $x \in U$, il existe $r > 0$ tel que $B(x, r) \subset U$.
- On dit qu'un ensemble dans un espace métrique X est fermé si son complémentaire dans X est ouvert.

Définition. 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.

Exemple.

- Boréliens sur $\mathbb{R}$, sur $]0, 1[$.
- Boréliens sur $\mathbb{R}^2$.

Espace mesurable

Définition. Soit Ω un ensemble et soit $\mathcal{A}$ une tribu sur Ω. On dit que $(\Omega, \mathcal{A})$ est un espace mesurable.

Corollaire. *Quand on s'intéressera aux probabilités, on dira que $(\Omega, \mathcal{A})$ est un espace probabilisable.*

Propriété. *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 $A_1 \times \cdots \times A_n$ où chaque $A_i \in \mathcal{A}_i$. L'ensemble des ensembles élémentaires est une algèbre et on note $\mathcal{A}_1 \otimes \cdots \otimes \mathcal{A}_n$ (on dit $\mathcal{A}_1$ tensoriel $\mathcal{A}_2$ ... tensoriel $\mathcal{A}_n$) la tribu sur Ω engendrée par ces ensembles élémentaires.*

Exemple.

Pavés de $\mathbb{R}^d$.

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

Exemple.

Pile / Face 2 fois.

Définitions et Propriétés d'une mesure

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

- $\mu(\emptyset) = 0$.
- Pour tout $I \subset \mathbb{N}$ 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).

Définition. 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.

Exemple.

Cas de $\Omega = \mathbb{R}$, de $\mathbb{N}$, ou $\mathbb{R}^2$.

Définition. 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).

Remarque.

Sur $(\Omega, \mathcal{A})$, on peut définir une infinité de mesures.

Propriété. *Soit $(\Omega, \mathcal{A}, \mu)$ un espace mesuré et $(A_i)_{i\in\mathbb{N}}$, 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 $I \subset \mathbb{N}$, 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 $i \in \mathbb{N}$ (suite croissante en sens de l'inclusion), alors $(\mu(A_n))_{n\in\mathbb{N}}$ est une suite croissante convergente telle que $\mu\left(\bigcup_{i\in\mathbb{N}} A_i\right) = \lim_{i\to+\infty} \mu(A_i)$ (même si cette limite est $+\infty$).*
5. *Si $A_{i+1} \subset A_i$ pour tout $i \in \mathbb{N}$ (suite décroissante en sens de l'inclusion) et $\mu(A_0) < +\infty$, alors $(\mu(A_n))_{n\in\mathbb{N}}$ est une suite décroissante convergente telle que $\mu\left(\bigcap_{i\in\mathbb{N}} A_i\right) = \lim_{i\to+\infty} \mu(A_i)$.*

Exemple.

1. Soit $(\Omega, \mathcal{A}, \mu)$ un espace mesuré. On définit $\nu(A) = \mu(A \cap B)$ où $B \in \mathcal{A}$. ν mesure ?
2. Si μ_1 et μ_2 mesures sur $(\Omega, \mathcal{A})$, $\mu_1 + \mu_2$ et $\alpha\mu$ sont-elles des mesures ?

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

1. On définit $\limsup(A_n)_n = \bigcap_{n\in\mathbb{N}} \bigcup_{m\geq n} A_m$ (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 $\liminf(A_n)_n = \bigcup_{n\in\mathbb{N}} \bigcap_{m\geq n} A_m$ (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).

Exemple.

Cas des suites croissantes et décroissantes d'ensembles.

Théorème (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 $\mathcal{A}$.*

Exemple.

Définition de la mesure de Lebesgue sur $\mathbb{R}$, $\mathbb{R}^n$,...

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

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

Exemple.

- Mesure de Lebesgue sur $\mathbb{N}$ ou $\mathbb{Q}$.
- La propriété " la suite de fonction $f_n(x) = x^n$ converge vers la fonction $f(x) = 0$" est vraie λ-presque partout sur $[0,1]$.
- Soit $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mu)$ et soit F la fonction définie par $F(x) = \mu(]-\infty, x])$ pour $x\in\mathbb{R}$.

Fonctions mesurables

Rappel. 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).

Définition. Soit $f : E\mapsto F$ et soit $\mathcal{I}$ une tribu sur F. On note $f^{-1}(\mathcal{I})$ l'ensemble de sous-ensembles de Ω tel que $f^{-1}(\mathcal{I}) = \{f^{-1}(I), I\in\mathcal{I}\}$.

Propriété. *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. 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 $f^{-1}(A') \in \mathcal{A}$).

Exemple.

- Fonction indicatrice.
- Combinaison linéaire de fonctions indicatrices.

Remarque.

Dans le cas où $(\Omega, \mathcal{A})$ est un espace probabilisable, et si $f : \Omega \mapsto \mathbb{R}$, alors si f est une fonction mesurable sur $\mathcal{A}$ et $\mathcal{B}(\mathbb{R})$, alors f est une variable aléatoire.

Exemple.

Nombre de Piles dans un jeu de Pile/Face.

Remarque.

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 de Ω', si f est une fonction mesurable sur $\mathcal{A}$ et $\mathcal{B}(\Omega')$, alors f est dite fonction borélienne.

Proposition. *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) $\Longleftrightarrow$ $(f^{-1}(\mathcal{F}) \subset \mathcal{A})$*

Conséquence.

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

Propriété.

- *Soit f mesurable de $(\Omega, \mathcal{A})$ dans $(\Omega', \mathcal{A}')$ et g mesurable de $(\Omega', \mathcal{A}')$ dans $(\Omega'', \mathcal{A}'')$. Alors $g_0 f$ est mesurable dans $\mathcal{A}$ et $\mathcal{A}'$.*
- *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 dans $\mathcal{A}$ et $\mathcal{A}_1 \otimes \mathcal{A}_2$.*
- *Soit $(f_n)_{n \in \mathbb{N}}$ 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 dans $\mathcal{A}$ et $\mathcal{B}(\Omega')$.*

Définition. 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.

Cas particulier.

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

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

Définition. Soit $f : \Omega \to \mathbb{R}$. 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 $f(\omega) = \sum_{i=1}^{n} \alpha_i \mathbb{I}_{A_i}(\omega)$.

Remarque.

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

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

Conséquence. Soit f une fonction réelle mesurable. Alors f est limite simple de fonctions étagées.

Intégration de Lebesgue

Dans toute la suite, on considère $(\Omega, \mathcal{A}, \mu)$ un espace mesuré.

Intégrale de Lebesgue d'une fonction positive

Définition.

1. Soit $f = \mathbb{I}_a$, où $A \in \mathcal{A}$. Alors :
$$\int f\, d\mu = \int_\omega f(\omega) d\mu(\omega) = \mu(A).$$
2. Soit $f = \mathbb{I}_a$, où $A \in \mathcal{A}$ et soit $B \in \mathcal{A}$. Alors :
$$\int_B f\, d\mu = \int_B f(\omega) d\mu(\omega) = \int \mathbb{I}_B \mu(A)(\omega) f(\omega) d\mu(\omega) = \mu(A \cap B).$$
3. Soit f une fonction étagée positive telle que $f = \sum_{i=1}^{n} \alpha_i \mathbb{I}_{A_i}$, où les $A_i \in \mathcal{A}$ et $\alpha_i > 0$ et soit $B \in \mathcal{A}$. Alors :
$$\int_B f\, d\mu = \int_B f(\omega) d\mu(\omega) = \int \mathbb{I}_B(\omega) f(\omega) d\mu(\omega) = \sum_{i=1}^{n} \alpha_i \mu(A_i \cap B).$$

Exemple.

Fonction $\mathbb{I}_{\mathbb{Q}}$, fonctions en escalier,...

Définition. 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 :

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

Propriété. *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 une fonction* $\mathcal{A}$*-mesurable positive 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 (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 :*

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

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

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

Exemple.

Appliquer Fatou à (f_n) telle que $f_{2n} = \mathbb{I}_A$ et $f_{2n+1} = \mathbb{I}_B$.

Intégrale de Lebesgue d'une fonction réelle et propriétés

Définition. Soit $(\Omega, \mathcal{A}, \mu)$ un espace mesuré, $B \in \mathcal{A}$ et soit f une fonction $\mathcal{A}$-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

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

Notation. Lorsque f est μ-intégrable sur B, soit $\int |f|\, d\mu < +\infty$, on note $f \in \mathcal{L}^1(\Omega, \mathcal{A}, \mu)$ (on dit que f est $\mathcal{L}^1$).

Exemple.

Intégrale de Riemann et intégrale de Lebesgue.
Cas de la masse de Dirac.

Propriété. *On suppose que* 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* $(\alpha, \beta) \in \mathbb{R}^2$.
2. *Si* $f \leq g$ *alors* $\int f\, d\mu \leq \int g\, d\mu$.

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

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

Extension.

Le Théorème de Lebesgue s'applique également dans le cas où $(f_n)_n$ converge presque partout vers f.

Exemple.

Convergence d'intégrale dépendant d'un paramètre : par exemple $\displaystyle\int_0^\infty \frac{f(x)}{1+x^n} dx$.

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

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

Exemple.

Soit X une v.a. sur $(\Omega, \mathcal{A}, \mathbb{P})$. Alors $\phi\left(\mathbb{E}[X]\right) \leq \mathbb{E}\left(\phi(X)\right)$.

Mesures induites et densités

Théorème (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 $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Alors, si $\phi_0 f \in \mathcal{L}^1(\Omega, \mathcal{A}, \mu)$,*

$$\int_{\Omega'} \phi \, d\mu_f = \int_\Omega \phi_0 f \, d\mu.$$

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

Propriété. *Soit $(\Omega, \mathcal{A}, \mu)$ un espace mesuré et f une fonction définie sur $(\Omega, \mathcal{A})$ mesurable et positive. On suppose que pour $A \in \mathcal{A}$, $\nu(A) = \displaystyle\int_A f \, d\mu$. 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. On dit que μ mesure sur $(\Omega, \mathcal{A})$ est σ-finie 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 (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 densité de ν par rapport à μ, telle que pour tout $A \in \mathcal{A}$, $\nu(A) = \displaystyle\int_A f \, d\mu$.*

Théorème (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 $f : \Omega \mapsto \mathbb{R}$, $\mathcal{A}$-mesurable et μ-intégrable. alors :*

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

Espaces $\mathcal{L}^p$

Définition. 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 $f : \Omega \mapsto \mathbb{R}$, mesurables et telles que $\int |f|^p d\mu < +\infty$.

Définition. Pour $f \in \mathcal{L}^p(\Omega, \mathcal{A}, \mu)$, où $p > 0$, on note $\| f \|_p = \left(\int |f|^p d\mu \right)^{1/p}$.

Propriété (Inégalité de Hölder). *Soit $p > 1$ et $q > 1$ 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 (Inégalité de Minkowski). *Soit $p > 1$ 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*

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

Remarque.

Pour $p > 1$, $\| \,.\, \|_p$ définie ainsi sur une semi-norme sur $\mathcal{L}^p(\Omega, \mathcal{A}, \mu)$. Pour obtenir une norme, il faut se placer dans l'espace $\mathbb{L}^p(\Omega, \mathcal{A}, \mu)$ 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 $\mathbb{L}^p(\Omega, \mathcal{A}, \mu)$ on dira que $f = g$ lorsque $f = g$ μ-presque partout).

Définition. Pour f et $g \in \mathbb{L}^2(\Omega, \mathcal{A}, \mu)$, on définit le produit scalaire $< f, g >= \int f.g\, d\mu$. On muni ainsi $\mathbb{L}^2(\Omega, \mathcal{A}, \mu)$ d'une structure d'espace de Hilbert. On dira que f est orthogonale à g lorsque $< f, g >= 0$.

Conséquence. Si A est un sous-espace vectoriel fermé de $\mathbb{L}^2(\Omega, \mathcal{A}, \mu)$ (par exemple un sous-espace de dimension finie), alors pour tout $f \in \mathbb{L}^2(\Omega, \mathcal{A}, \mu)$, il existe un unique projeté orthogonal de f sur A, noté f_A, qui vérifie $f_A = \underset{g \in A}{\text{Arginf}} \| g - f \|_2$.

1 Variables aléatoires et espérance

1.1 Variables aléatoires

Définition. On dit que X est une variable aléatoire sur $(\Omega, \mathcal{A}, \mathbb{P})$ un espace de probabilité si: $X : \Omega \to \mathbb{R}$ est une application mesurable de $(\Omega, \mathcal{A})$ dans $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$.

Pourquoi demander à X d'être mesurable? Parce que l'on veut pouvoir définir une fonction de répartition pour X, soit $F_X(x) = \mathbb{P}(\text{"}X \leq x\text{"}) = \mathbb{P}\big(\{\omega \in \Omega,\ X(\omega) \leq x\}\big) = X^{-1}(]-\infty, x])$: il faut donc que l'ensemble $X^{-1}(]-\infty, x])$ soit un évenement de $\mathcal{A}$ pour tout $x \in \mathbb{R}$.

Dans la suite, pour une variable aléatoire X définie sur $(\Omega, \mathcal{A}, \mathbb{P})$, on n'utilisera presque jamais la forme explicite de la fonction $\omega \to X(\omega)$ qui restera inconnue. En revanche, on préfèrera travailler avec la **loi** de X.
Qu'appelle-t-on loi de X? Il y a en réalité plusieurs moyens de la définir:

1. On définit la loi par $F_X(x) = \mathbb{P}(X \leq x)$ pour $x \in \mathbb{R}$, la **fonction de répartition** de X;
2. On définit la loi par la **mesure de probabilité $\mathbb{P}_X$ induite par** X: pour $B \in \mathcal{B}(\mathbb{R})$, $\mathbb{P}_X(B) = \mathbb{P}(X \in B) = \mathbb{P}\big(X^{-1}(B)\big)$. On notera que $\mathbb{P}_X$ est une mesure de probabilité définie sur $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ ou bien sur $(X(\Omega), \mathcal{T})$, avec $\mathcal{T}$ une tribu sur $X(\Omega) \subset \mathbb{R}$.

3. Dans 2 cas particuliers, on préfèrera travailler avec:

   - Si $X(\Omega) = \{x_i\}_{i\in I}$ avec $I \subset \mathbb{N}$ et $x_i \in \mathbb{R}$ for all $i \in I$, on parlera de variable aléatoire **discrête** et la loi sera donnée par les $\mathbb{P}(X = x_i) = \mathbb{P}_X(\{x_i\})$ pour $i \in I$. On remarque alors que pour tout $B \in \mathcal{B}(\mathbb{R})$, $\mathbb{P}_X(B) = \sum_{i\in I} \mathbb{P}(X = x_i)\, \delta_{\{x_i\}}(B)$, avec δ masse (mesure) de Dirac.
   - Si $X(\Omega)$ est une union finie ou dénombrable d'intervalles de I, et si $\mathbb{P}_X$ est absolument continue par rapport à la mesure de Lebesgue sur $\mathbb{R}$, alors on peut définir la **densité de probabilité** f_X de X par rapport à la mesure de Lebesgue sur $\mathbb{R}$, et l'on a:

     $$F_X(x) = \int_{-\infty}^{x} f_X(t)\, dt \quad \text{pour tout } x \in \mathbb{R}.$$

     La fonction F_X est une fonction **absolument continue** sur $\mathbb{R}$, elle est même dérivable λ-presque partout sur $\mathbb{R}$ et $F'_X(x) = f_X(x)$ pour presque tout x dans $\mathbb{R}$. On dira souvent pour simplifier que X est une variable aléatoire **"absolument continue"** et même parfois que X est une variable aléatoire **"continue"**.

4. La loi est donnée par fonction caractéristique de X soit $\phi_X(u) = \mathbb{E}\left[e^{i\,u\,X}\right]$ pour $u \in \mathbb{R}$ (voir plus loin).

Remarque.

On peut connaître le "type" de variable aléatoire qu'est la variable X grâce au graphe de la fonction de répartition F_X:

1. Si F_X est une fonction en escalier avec un nombre fini ou dénombrable de sauts en les $(x_i)_{i\in I}$ où $I \subset \mathbb{N}$, alors X est une variable discrète de loi $\mathbb{P}(X = x_i) = F_X(x_i) - \lim_{x\to x_i^-} F_X(x)$.

2. Si F_X est une fonction continue sur $\mathbb{R}$ et dérivable sauf en un nombre fini ou dénombrable de points, alors X est une variable (absolument) continue, et la dérivée de F_X lorsqu'elle existe est la densité de probabilité f_X. Pour x_0 tel que F_X n'est pas dérivable en x_0, on pourra choisir que $f_X(x_0) = 0$ ou tout autre réel.

3. Si aucun des 2 cas précédents n'est vérifié, X pourra être un "mélange" de loi discrète et continue, soit $\mathbb{P}_X(A) = \sum_{i\in I} p_i\, \delta_{x_i}(A) + \int_A g(x) d\lambda(x)$, ou bien une loi diffuse non absolument continue par rapport à la mesure de Lebesgue comme la loi de Cantor.

Dans la propriété suivante, on va utiliser pleinement le fait qu'une variable aléatoire est une fonction mesurable:

Propriété. *Soit X et Y deux v.a. sur $(\Omega, \mathcal{A}, \mathbb{P})$ et $g : \mathbb{R}^2 \to \mathbb{R}$ une fonction borélienne. Alors $Z = g(X, Y)$ est une v.a. sur $(\Omega, \mathcal{A}, \mathbb{P})$.*

Proof. Voir cours Intégration et Probabilités: il s'agit de montrer que la composition d'une fonction mesurable par une autre fonction mesurable est une fonction mesurable. Et également que le vecteur composé de fonctions mesurables est une fonction mesurable. □

Par itération du procédé, la propriété est aussi vraie pour une fonction g à n variables. Mais aussi bien-sûr pour une fonction à 1 variable.
De la même manière, toujours du fait qu'une variable aléatoire est une fonction mesurable:

Propriété. *Soit X et Y deux v.a. sur $(\Omega, \mathcal{A}, \mathbb{P})$ telles que $\mathbb{P}_X = \mathbb{P}_Y$, ce que l'on peut aussi noter $X \overset{\mathcal{L}}{\sim} Y$. Alors pour et $g : \mathbb{R} \to \mathbb{R}$ une fonction borélienne, $g(X) \overset{\mathcal{L}}{\sim} g(Y)$.*

Proof. Pour tout $B \in \mathcal{B}(\mathbb{R})$, $\mathbb{P}_{g(X)}(B) = \mathbb{P}(g(X) \in B) = \mathbb{P}(X \in g^{-1}(B)) = \mathbb{P}_X(g^{-1}(B))$. Mais on a aussi $\mathbb{P}_{g(Y)}(B) = \mathbb{P}_Y(g^{-1}(B)) = \mathbb{P}_X(g^{-1}(B))$ puisque $\mathbb{P}_X = \mathbb{P}_Y$. D'où le résultat. □

Remarque: Si $X \overset{\mathcal{L}}{\sim} \mathcal{U}\big([0,1]\big)$ et $Y = 1 - X$, alors on a aussi $Y \overset{\mathcal{L}}{\sim} \mathcal{U}\big([0,1]\big)$, donc $\mathbb{P}_X = \mathbb{P}_Y$ (on le montre en utilisant par exemple la fonction de répartition). Mais pourtant on n'a pas du tout $X = Y$...

Définition. Si X est une v.a. sur $(\Omega, \mathcal{A}, \mathbb{P})$ et $0 < p < 1$, le quantile d'ordre p de X est : $q_X(p) = \inf\big\{y \in \mathbb{R},\ F_X(y) \geq p\big\}$.

Cas particulier: Si $p = 1/2$, alors $q_X(p)$ est la médiane (théorique) de X.

Propriété. *Si X est une v.a. continue, $q_X(p) = \widetilde{F}_X^{-1}(p)$ où $\widetilde{F}_X : x \in X(\Omega) \mapsto F_X(x)$.*

Exemple: Si X suit une distribution $\mathcal{E}(\lambda)$ (exponentielle),

$$F_X(x) = (1 - e^{-\lambda x})\mathbb{I}_{x \geq 0} \quad \text{and} \quad q_X(1/2) = \frac{\ln(2)}{\lambda} \neq \mathbb{E}[X] = \frac{1}{\lambda}.$$

1.2 Espérance de variables aléatoires

Définition. Soit X une variable aléatoire sur $(\Omega, \mathcal{A}, \mathbb{P})$ un espace probabilisé. Alors si $X \in \mathbb{L}^1(\Omega, \mathcal{A}, \mathbb{P})$ (donc si $\int_\Omega |X(\omega)|\, d\mathbb{P}(\omega) < \infty$), on définit l'espérance de X par le réel:

$$\mathbb{E}[X] = \int X d\mathbb{P} = \int_\Omega X(\omega)\, d\mathbb{P}(\omega).$$

Plus généralement, si $\phi : \mathbb{R} \mapsto \mathbb{R}$ est borélienne et si $\phi(X) \in \mathbb{L}^1(\Omega, \mathcal{A}, \mathbb{P})$, on définit l'espérance de $\phi(X)$ par

$$\mathbb{E}[\phi(X)] = \int \phi(X) d\mathbb{P} = \int_\Omega \phi(X(\omega))\, d\mathbb{P}(\omega).$$

Propriété. *Si X est une variable aléatoire sur $(\Omega, \mathcal{A}, \mathbb{P})$, si $\phi : \mathbb{R} \mapsto \mathbb{R}$ est borélienne telle que $\phi(X) \in \mathbb{L}^1(\Omega, \mathcal{A}, \mathbb{P})$, alors :*

$$\mathbb{E}\left[\phi(X)\right] = \int_{\mathbb{R}} \phi(x)\, d\mathbb{P}_X(x).$$

Proof. Théorème du transport... □

Conséquence.

- Si $\mathbb{P}_X$ est absolument continue par rapport à la mesure de Lebesgue (donc X est une v.a. dite absolument continue), de densité f_X, alors $\mathbb{E}[\phi(X)] = \displaystyle\int_{\mathbb{R}} \phi(x)\, f_X(x)\, dx$.
- Si $\mathbb{P}_X$ est absolument continue par rapport à une mesure de comptage sur $\{x_i\}_{i \in I}$ avec $I \subset \mathbb{N}$ (donc X est une v.a. dite discrète), de densité p_X avec $p_X(i) = \mathbb{P}(X = x_i)$, alors $\mathbb{E}[\phi(X)] = \displaystyle\sum_{i \in I} p_X(i)\, \phi(x_i)$.

Propriété.

1. *Soit X et Y des variables aléatoires telles que X et $Y \in \mathbb{L}^1(\Omega, \mathcal{A}, \mathbb{P})$. Alors pour tout $(a, b) \in \mathbb{R}^2$, $aX + bY \in \mathbb{L}^1(\Omega, \mathcal{A}, \mathbb{P})$ et*

$$\mathbb{E}[a\,X + b\,Y] = a\,\mathbb{E}[X] + b\,\mathbb{E}[Y].$$

2. *Soit X une variable aléatoire sur $(\Omega, \mathcal{A}, \mathbb{P})$, et soit $B \in \mathcal{B}(\mathbb{R})$. Alors $\mathbb{E}[\mathbb{I}_B(X)] = \mathbb{P}(X \in B)$.*

3. *Si X est une variable aléatoire sur $(\Omega, \mathcal{A}, \mathbb{P})$, si $\phi : \mathbb{R} \mapsto \mathbb{R}$ est une fonction borélienne* **convexe** *telle que X et $\phi(X) \in \mathbb{L}^1(\Omega, \mathcal{A}, \mathbb{P})$, alors*

$$\mathbb{E}[\phi(X)] \geq \phi(\mathbb{E}[X]) \quad \textit{(Inégalité de Jensen).}$$

4. *Soit X et Y des variables aléatoires telles que $X \in \mathbb{L}^p(\Omega, \mathcal{A}, \mathbb{P})$ et $Y \in \mathbb{L}^q(\Omega, \mathcal{A}, \mathbb{P})$ avec $\frac{1}{p} + \frac{1}{q} = 1$ où $p > 1$, $q > 1$. Alors $XY \in \mathbb{L}^1(\Omega, \mathcal{A}, \mathbb{P})$ et*

$$\mathbb{E}[|XY|] \leq (\mathbb{E}[|X|^p])^{1/p} (\mathbb{E}[|Y|^q])^{1/q} \quad \textit{(Inégalité de Hölder).}$$

5. *Soit X et Y des variables aléatoires telles que X et $Y \in \mathbb{L}^p(\Omega, \mathcal{A}, \mathbb{P})$, avec $p \geq 1$. Alors $X + Y \in \mathbb{L}^p(\Omega, \mathcal{A}, \mathbb{P})$ et*

$$(\mathbb{E}[|X+Y|^p])^{1/p} \leq (\mathbb{E}[|X|^p])^{1/p} + (\mathbb{E}[|Y|]^p)^{1/p} \quad \textit{(Inégalité triangulaire de Minkowski).}$$

Proof.

1. On a d'abord $aX + bY$ qui est une v.a., puis $\mathbb{E}[|aX + bY|] \leq |a|\mathbb{E}[|X|] + |b|\mathbb{E}[|Y|] < \infty$. On utilise ensuite la linéarité de l'intégrale.
2. La v.a. $\mathbb{I}_B(X)$ est une v.a. de Bernoulli de paramètre $\mathbb{P}(X \in B)$, et donc également d'espérance $\mathbb{P}(X \in B)$.
3. Vue en "Intégration et probabilités".
4. Plus généralement, on a $\int |fg| d\mu \leq \|f\|_p \|g\|_q$ avec $\|f\|_p = \left(\int |f|^p d\mu\right)^{1/p}$. En effet, la fonction $x \in]0, \infty[\mapsto -\ln(x)$ est convexe, donc pour tout $a > 0$, $b > 0$ et $\theta \in [0, 1]$, $-\log(a\theta + b(1-\theta)) \leq -\theta \log(a) - (1-\theta)\log b$. En passant à l'exponentielle, on en déduit que $a^\theta b^{1-\theta} \leq a\theta + b(1-\theta)$.
   Pour tout $x \in \Omega$, choisissons $a = \frac{|f(x)|^p}{\|f\|_p^p}$, $b = \frac{|g(x)|^q}{\|f\|_q^q}$ et $\theta = 1/p$. Alors pour tout $x \in \Omega$ et avec $1 - \theta = 1/q$,
   $$\frac{|f(x)|}{\|f\|_p} \frac{|g(x)|}{\|g\|_q} \leq \frac{1}{p} \frac{|f(x)|^p}{\|f\|_p^p} + \frac{1}{q} \frac{|g(x)|^q}{\|g\|_q^q}.$$
   Intégrons des 2 côtés (on peut et on garde le sens de l'inégalité), on obtient:
   $$\frac{\int |fg| d\mu}{\|f\|_p \|g\|_q} \leq \frac{1}{p} \frac{\int |f|^p d\mu}{\|f\|_p^p} + \frac{1}{q} \frac{\int |g|^q d\mu}{\|g\|_q^q} \leq \frac{1}{p} + \frac{1}{q} = 1,$$
   d'où le résultat.
5. On va montrer plus généralement que $\|f + g\|_p \leq \|f\|_p + \|g\|_p$ pour $p \geq 1$. Il est clair que si $\|f + g\|_p = 0$, le résultat est vrai. De même si $p = 1$. Pour $p > 1$,
   $$\begin{aligned}
   \|f+g\|_p^p &\leq \int (|f| + |g|)|f+g|^{p-1} d\mu \\
   &\leq \int |f|\,|f+g|^{p-1} d\mu + \int |g|\,|f+g|^{p-1} d\mu \\
   &\leq \|f\|_p \left(\int |f+g|^{q(p-1)} d\mu\right)^{1/q} + \|g\|_p \left(\int |f+g|^{q(p-1)} d\mu\right)^{1/q} \quad \text{(Inégalité de Hölder)} \\
   &\leq \left(\|f\|_p + \|g\|_p\right) \|f+g\|_p^{p-1},
   \end{aligned}$$
   d'où l'inégalité.

□

Conséquence. Soit X une variable aléatoire telle que $X \in \mathbb{L}^p(\Omega, \mathcal{A}, \mathbb{P})$ pour $p \geq 1$, c'est-à-dire l'ensemble des variables aléatoires définies à la classe d'équivalence $\mathbb{P}$-p.s. près (donc $X \sim Y$ si $X - Y = 0$ $\mathbb{P} - p.s.$) et telle que $\mathbb{E}[|X|^p] < \infty$. Alors $\|X\|_p = (\mathbb{E}[|X|^p])^{1/p}$ est une norme sur $\mathbb{L}^p(\Omega, \mathcal{A}, \mathbb{P})$.

Proof. Pour que ce soit bien une norme, il faut que:

1. $\|X\|_p \geq 0$ pour tout X, ce qui est évidemment vrai;
2. pour tout $\lambda \in \mathbb{R}$, $\|\lambda X\|_p = |\lambda|\,\|X\|_p$ ce qui est vrai également;
3. l'inégalité triangulaire soit vérifiée, soit $\|X + Y\|_p \leq \|X\|_p + \|Y\|_p$, ce qui est vrai grâce à l'Inégalité de Minkowski;

4. si $\|X\|_p = 0$, soit $\int_\Omega |X(\omega)|\, d\mathbb{P}(\omega) = 0$ alors $X = 0$, ce qui est vrai si l'on considère $X = 0$ $\mathbb{P}$-p.s.

La quatrième propriété implique de travailler sur $\mathbb{L}^p(\Omega, \mathcal{A}, \mathbb{P})$ et non sur $\mathcal{L}^p(\Omega, \mathcal{A}, \mathbb{P})$ qui est constitué des fonctions f telles que $\|f\|_p < \infty$ sans la classe d'équivalence: sur $\mathcal{L}^p(\Omega, \mathcal{A}, \mathbb{P})$ on a seulement une quasi-norme, la propriété 4/ n'est pas vérifiée (prendre par exemple $\Omega = [0,1]$, $\mathcal{A} = \mathcal{B}([0,1])$, $\mathbb{P} = \lambda$ et $f(x) = 0$ pour $x \in]0,1]$ et $f(0) = 1$). □

Conséquence. Soit X une variable aléatoire telle que $X \in \mathbb{L}^p(\Omega, \mathcal{A}, \mathbb{P})$ pour $p > 0$. Alors pour tout $1 \le r \le p$, $X \in \mathbb{L}^r(\Omega, \mathcal{A}, \mathbb{P})$ et

$$(\mathbb{E}[|X|^r])^{1/r} \le (\mathbb{E}[|X|^p])^{1/p} \Longleftrightarrow \|X\|_r \le \|X\|_p.$$

Proof. La fonction $\phi(x) = x^{p/r}$ est convexe sur $[0, \infty[$ car $p/r \ge 1$ (il suffit de dériver 2 fois et $\phi'' \ge 0$). On applique alors l'Inégalité de Jensen à la variable $|X|^r$: $\mathbb{E}[\phi(|X|^r)] \ge \mathbb{E}[|X|^r]^{p/r}$, d'où le résultat. □

Définition. Pour X et Y des variables aléatoires telles que X et $Y \in \mathbb{L}^2(\Omega, \mathcal{A}, \mathbb{P})$, on définit:

- la variance de X, $\mathrm{var}(X) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$.
- la covariance de X et Y par

$$\mathrm{cov}(X, Y) = \mathbb{E}\left[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\right] = \mathbb{E}[X\,Y] - \mathbb{E}[X]\,\mathbb{E}[Y].$$

Remarque: On a $\mathrm{var}(X) = \mathrm{cov}(X, X)$.

Propriété. *Sur l'espace vectoriel* $E = \left\{X \in \mathbb{L}^2(\Omega, \mathcal{A}, \mathbb{P}),\ \mathbb{E}[X] = 0\right\}$, *on définit* $< X, Y >= cov(X, Y)$ *pour* $X, Y \in E$. *Alors* $< \cdot, \cdot >$ *définit un produit scalaire.*

Conséquence. Pour X et Y deux v.a. sur $\mathbb{L}^2(\Omega, \mathcal{A}, \mathbb{P})$, alors

$$|\mathrm{cov}(X, Y)|^2 \le \mathrm{var}(X)\,\mathrm{var}(Y).$$

Proof. Soit X' et Y' telles que $\mathbb{E}[X'] = \mathbb{E}[Y'] = 0$: il suffit alors d'appliquer l'inégalité de Cauchy-Schwarz (cas particulier de Hölder avec $p = q = 2$): $|\mathbb{E}[X'\,Y']| \le \mathbb{E}[|X'\,Y'|] \le \sqrt{\mathbb{E}[X'^2]\,\mathbb{E}[Y'^2]}$. Ensuite on remplace X' par $X - \mathbb{E}[X]$ et Y' par $Y - \mathbb{E}[Y]$. □

Propriété. *Pour* X *et* Y *deux v.a. sur* $\mathbb{L}^2(\Omega, \mathcal{A}, \mathbb{P})$, *on peut définir le coefficient de corrélation entre* X *et* Y *par:*

$$cor(X, Y) = \frac{cov(X, Y)}{\sqrt{var(X)\,var(Y)}} \quad et \quad \left|cor(X, Y)\right| \le 1.$$

1.3 Fonction génératrice et fonction caractéristique

Définition. Si X est une v.a. sur $(\Omega, \mathcal{A}, \mathbb{P})$ à valeurs dans $\mathbb{N}$, on définit la fonction génératrice de X par :

$$g(z) = \mathbb{E}\left[z^X\right] \quad \text{pour tout } z \in [-1, 1].$$

Remarque: $g(z) = \sum_{k=0}^{\infty} \mathbb{P}(X = k)\, z^k$ est une série entière de rayon de convergence $R \ge 1$.

Propriété. *La fonction* g *est une fonction* $\mathcal{C}^\infty((-1, 1))$ *et* $\mathbb{P}(X = k) = g^{(k)}(0)/k!$.

Proof. Fonction $\mathcal{C}^\infty((-1,1))$: propriété d'une série entière. De plus, par unicité du développement en série entière on a également $g(z) = \sum_{k=0}^{\infty} \frac{g^{(k)}(0)}{k!}\, z^k = \sum_{k=0}^{\infty} \mathbb{P}(X = k)\, z^k$, d'où le résultat. □

Conséquence: $(X \overset{\mathcal{L}}{\sim} Y) \Longleftrightarrow (g_X = g_Y)$.

Propriété. *Si* $\mathbb{E}[|X|] < \infty$, *la fonction* g *est telle que* $g'(1) = \mathbb{E}[X]$. *On peut également obtenir pour tout* $k \geq 1$, $\mathbb{E}\big[X(X-1) \times \cdots \times (X-k)\big] = g^{(k+1)}(1)$ *lorsque* $\mathbb{E}[X^{k+1}] < \infty$.

Proof. Pour $|z| < 1$, $g'(z) = \sum_{j=1}^{\infty} \mathbb{P}(X = j)\, j\, z^{j-1}$. L'hypothèse $\mathbb{E}[X] < \infty$ implique que cette série entière est normalement convergente sur $[-1, 1]$ et ainsi $g'(1) = \mathbb{E}[X]$.
De plus, pour $|z| < 1$, $g^{(k+1)}(z) = \sum_{j=k}^{\infty} \mathbb{P}(X = j)\, j(j-1) \times \cdots \times (j-k)\, z^{j-k}$. Donc si $\mathbb{E}[X^{k+1}] < \infty$ pour $k \geq 1$, alors $\mathbb{E}\big[X(X-1) \times \cdots \times (X-k)\big] < \infty$ et $\mathbb{E}\big[X(X-1) \times \cdots \times (X-k)\big] = \sum_{j=k}^{\infty} \mathbb{P}(X = j)\, j(j-1) \times \cdots \times (j-k) = g^{(k+1)}(1)$. □

Exemple: Si $X \sim \mathcal{P}(\lambda)$ (loi de Poisson) avec $\lambda > 0$. Alors $g(z) = e^{\lambda(z-1)}$ pour $z \in \mathbb{R}$.

A quoi sert la fonction génératrice?

- A remplacer une mesure de probabilité par une fonction (dans le cas des v.a. à valeurs entières positives) puisque la fonction génératrice caractérise la loi;
- Mais surtout, à obtenir la loi ou montrer la convergence de sommes de variables indépendantes.

Une extension de la fonction génératrice à toute variable aléatoire, discrète, continue, ou autre, est obtenue par la fonction caractéristique:

Définition. Si X est une v.a. sur $(\Omega, \mathcal{A}, \mathbb{P})$, on définit la **fonction caractéristique** de X par:

$$\phi_X(u) = \mathbb{E}\big[e^{i\,u\,X}\big] = \mathbb{E}[\cos(u\,X)] + i\,\mathbb{E}[\sin(u\,X)] \quad \text{pour tout } u \in \mathbb{R}.$$

Remarque: ϕ_X est défine sur $\mathbb{R}$ du fait que $|e^{i\,u\,X}| \leq 1$ pour tout $u \in \mathbb{R}$ et en appliquant ensuite le Théorème de convergence dominée de Lebesgue.

Théorème. $\phi_X = \phi_Y \iff X$ *et* Y *ont la même distribution de probabilité.*

Proof. On peut prouver que pour tout $a \leq b$ tels que $\mathbb{P}_X(\{a, b\}) = 0$, alors

$$\mathbb{P}(X \in [a, b]) = \lim_{T \to \infty} \frac{1}{2\,\pi} \int_{-T}^{T} \frac{e^{-ita} - e^{-itb}}{it}\, \phi_X(t)\, dt.$$

Pour commencer, on montre que pour $c \in \mathbb{R}$, $\int_{-\infty}^{\infty} \frac{\sin(c\,t)}{t}\, dt = \pi$ si $c > 0$ et $= -\pi$ si $c < 0$, $= 0$ si $c = 0$. Tous ces cas se ramène au calcul de $I(0) = \int_0^{\infty} \frac{\sin(t)}{t}\, dt$. Cette intégrale existe en tant qu'intégrale de Riemann (mais pas en tant qu'intégrale de Lebesgue!). En effet, la fonction $t \to \frac{\sin(t)}{t}$ est prolongeable par continuité en 0 et en $+\infty$, on utilise une intégration par parties

$$\int_1^T \frac{\sin(t)}{t}\, dt = \big[-\frac{\cos(t)}{t}\big]_1^T - \int_1^T \frac{\cos(t)}{t^2}\, dt = \cos(1) - \int_1^T \frac{\cos(t)}{t^2}\, dt.$$

Cette dernière intégrale existe car la fonction est majorée par $\frac{1}{t^2}$ qui est intégrable en $+\infty$.
Par ailleurs, on considère pour $t \geq 0$, $I(x) = \int_0^{\infty} \frac{\sin(t)}{t}\, e^{-xt}\, dt$. Cette fonction existe et pour $x > 0$ elle est dérivable et $I'(x) = -\int_0^{\infty} \sin(t)\, e^{-xt}\, dt$ pour $x > 0$. Or

$$\int_0^{\infty} \sin(t)\, e^{-xt}\, dt = \int_0^{\infty} \mathcal{I}m(e^{it})\, e^{-xt}\, dt = \mathcal{I}m\Big(\int_0^{\infty} e^{it-xt}\, dt\Big) = \mathcal{I}m\Big(-\frac{1}{i-x}\Big) = \frac{1}{1+x^2}.$$

On en déduit que $I'(x) = -\frac{1}{1+x^2}$ soit $I(x) = -\arctan(x) + K$ où $K \in \mathbb{R}$. Comme on sait que $\lim_{x \to \infty} I(x) = 0$, donc $K = \pi/2$.
Il suffit de montrer que la fonction I est continue en 0, d'où $I(0) = \int_0^{\infty} \frac{\sin(t)}{t}\, dt = \pi/2$.

Revenons à la preuve et partons de $\lim_{T \to \infty} \frac{1}{2\,\pi} \int_{-T}^{T} \frac{e^{-ita} - e^{-itb}}{it}\, \phi_X(t)\, dt$. Soit:

$$\begin{aligned}
\frac{1}{2\,\pi} \int_{-T}^{T} \frac{e^{-ita} - e^{-itb}}{it}\, \phi_X(t)\, dt &= \frac{1}{2\,\pi} \int_{-T}^{T} \int_{-\infty}^{\infty} \frac{e^{it(x-a)} - e^{it(x-b)}}{it}\, d\mathbb{P}_X(x)\, dt \\
&= \frac{1}{2\,\pi} \int_{-\infty}^{\infty} \int_{-T}^{T} \frac{e^{it(x-a)} - e^{it(x-b)}}{it}\, dt\, d\mathbb{P}_X(x) \quad \text{(Fubini)} \\
&= \frac{1}{2\,\pi} \int_{-\infty}^{\infty} \int_{-T}^{T} \Big(\frac{\sin(t(x-a))}{t} - \frac{\sin(t(x-b))}{t}\Big)\, dt\, d\mathbb{P}_X(x) \quad \text{(Parité)}.
\end{aligned}$$

Mais si on s'intéresse à $\lim_{T\to\infty}\int_{-T}^{T}\left(\frac{\sin(t(x-a))}{t}-\frac{\sin(t(x-b))}{t}\right)dt=\int_{-\infty}^{\infty}\left(\frac{\sin(t(x-a))}{t}-\frac{\sin(t(x-b))}{t}\right)dt=\phi_{a,b}(x)$, alors $\phi_{a,b}(x)=0$ si $x>b$ ou $x<a$ et $\phi_{a,b}(x)=2\pi$ pour $a<x<b$. Pour $x=a$ et $x=b$, on obtient $\pm\pi$, mais de l'hypothèse $\mathbb{P}_X(\{a,b\})=0$ cela n'interviendra pas. D'où:

$$\lim_{T\to\infty}\frac{1}{2\,\pi}\int_{-T}^{T}\frac{e^{-ita}-e^{-itb}}{it}\,\phi_X(t)\,dt=\frac{1}{2\,\pi}\int_a^b 2\,\pi\,d\mathbb{P}_X(x)=\mathbb{P}(X\in[a,b]).$$

□

Remarque: Si X est une v.a. discrète à valeurs entières, alors $\phi_X(u)=g(e^{iu})$.

Propriété. *Si X est une v.a. définie sur $(\Omega,\mathcal{A},\mathbb{P})$ telle que $\mathbb{E}[|X|^k]<\infty$, alors ϕ_X est une fonction $\mathcal{C}^k(\mathbb{R})$ et $\phi_X^{(k)}(u)=i^k\,\mathbb{E}[X^k e^{iuX}]$ pour tout $u\in\mathbb{R}$.*

Proof. Pour $k=1$, $u\in\mathbb{R}$ et $h\neq 0$, nous avons $\frac{\phi_X(u+h)-\phi_X(u)}{h}=\mathbb{E}\left[e^{iuX}\left(\frac{e^{ihX}-1}{h}\right)\right]$. Il est clair que $e^{iuX}\left(\frac{e^{ihX}-1}{h}\right)\to i\,X\,e^{iuX}$ pour $h\to 0$ puisque $\frac{e^{ihx}-1}{h}\to i\,x$ lorsque $h\to 0$ pour tout $x\in\mathbb{R}$ (on peut considérer $g(h)=e^{ihx}$ et $(g(h)-g(0))/h\to g'(0)$ lorsque $h\to 0$). De plus, $\left|e^{iuX}\left(\frac{e^{iuh}-1}{h}\right)\right|\leq|X|$ pour tout $u\in\mathbb{R}$ et $h\neq 0$, et $\mathbb{E}[|X|]<\infty$ par hypothèse. Le théorème de Lebesgue implique alors que $\frac{\phi_X(u+h)-\phi_X(u)}{h}\to\mathbb{E}[iXe^{iuX}]=\phi'_X(u)$. Même type de preuve pour $k\geq 2$. □

Remarque: $\phi'_X(0)=i\,\mathbb{E}[X]$ et $\phi''_X(0)=-\mathbb{E}[X^2]$.

Voici une autre propriété de la fonction caractéristique, spécifique aux v.a. "continues":

Propriété. *Si X est une v.a. de loi continue par rapport à la mesure de Lebesgue, de densité f_X et de fonction caractéristique ϕ_X telle que $\int_{\mathbb{R}}|\phi_X(u)|\,du<\infty$, alors*

$$f_X(x)=\frac{1}{2\,\pi}\int_{-\infty}^{\infty}\phi_X(u)\,e^{-i\,u\,x}du\quad\textit{pour tout } x\in\mathbb{R}.$$

Proof. Dans le cas présent, on a: $\phi_X(u)=\int_{-\infty}^{\infty}f_X(x)\,e^{-i\,u\,x}du$. On peut reprendre la preuve d'identification de la loi mais en passant directement à la limite car la fonction caractéristique est alors de module intégrable. D'où pour tout $a<b$,

$$\mathbb{P}(X\in[a,b])=\int_a^b f_X(x)\,dx=\frac{1}{2\,\pi}\int_{-\infty}^{\infty}\frac{e^{-ita}-e^{-itb}}{it}\,\phi_X(t)\,dt=\frac{1}{2\,\pi}\int_{-\infty}^{\infty}\int_a^b e^{-itx}\,dx\,\phi_X(t)\,dt=\int_a^b\frac{1}{2\,\pi}\int_{-\infty}^{\infty}e^{-itx}\,\phi_X(t)\,dt\,dx,$$

par Fubini. D'où le résultat par identification puisqu'il est vrai pour tout $a<b$. □

A quoi sert la fonction caractéristique?

- Fonction caractéristique de la somme de v.a. indépendantes;
- Fonction caractéristique pour caractériser l'indépendance (voir un peu plus loin);
- Convergence d'une suite de fonctions caractéristiques vers une fonction caractéristique $\Longleftrightarrow$ convergence en loi (Théorème de Lévy, voir un peu plus loin).

2 Vecteurs aléatoires

2.1 Définitions et premières propriétés

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

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

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

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

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

Proof. On peut construire la fonction g_i qui associe à un vecteur de $\mathbb{R}^d$ sa ième coordonnée. La fonction sur $X_i : \Omega \to \mathbb{R}$ est donc telle que $X_i(\omega) = g_i(X(\omega))$ et comme g est continue, donc mesurable, X_i est une variable aléatoire.
Soit $(x_1^{(n)})_n, \ldots, (x_{i-1}^{(n)})_n, (x_{i+1}^{(n)})_n, \ldots, (x_d^{(n)})_n$ $d-1$ suites croissantes de réels telles que $x_j^{(n)} \underset{n\to\infty}{\longrightarrow} +\infty$ pour tout $j \neq i$. Il est clair que les pavés $A_n =]-\infty, x_1^{(n)}] \times \cdots \times]-\infty, x_{i-1}^{(n)}] \times]-\infty, x_i] \times]-\infty, x_{i+1}^{(n)}] \times \cdots \times]-\infty, x_d^{(n)}]$ forment une suite croissante de pavés de $\mathbb{R}^n$ (on a $A_m \subset A_{m+1}$ pour tout m). On sait alors que $\lim_{n\to\infty} \mathbb{P}_X(A_n) = \mathbb{P}_X(\bigcup_{n\in\mathbb{N}} A_n) = \mathbb{P}_X\big(\mathbb{R}^{i-1} \times]-\infty, x_i] \times \mathbb{R}^{d-i}\big) = \mathbb{P}(X_i \leq x_i) = F_{X_i}(x_i)$. Comme ceci est vrai pour toute suite croissante, on a bien le résultat. □

Remarque: Les mesures de probabilités $\mathbb{P}_{X_i}$ déterminées de façon univoque à partir des F_{X_i} sont appelées **lois marginales** de X. Cependant, les lois marginales ne permettent pas d'identifier la loi du vecteur, sauf à spécifier d'autres propriétés (indépendance par exemple).

Corollaire. *Si $F_X(x)$ est continue sur $\mathbb{R}^d$ et si $\frac{\partial^d}{\partial x_1 \cdots \partial x_d} F_X(x_1, \ldots, x_d)$ existe presque partout sur $\mathbb{R}^d$, alors la densité de la loi $\mathbb{P}_X$ par rapport à la mesure de Lebesgue sur $\mathbb{R}^d$ vaut:*

$$f_X(x_1, \ldots, x_d) = \frac{\partial^d}{\partial x_1 \cdots \partial x_d} F_X(x_1, \ldots, x_d) \quad \textit{pour presque tout } (x_1, \ldots, x_d) \in \mathbb{R}^d.$$

Propriété. *Soit $X = (X_1, \ldots, X_d)$ un vecteur aléatoire défini sur $(\Omega, \mathcal{A}, \mathbb{P})$ à valeurs dans $\mathbb{R}^d$ dont la mesure de probabilité est absolument continue par rapport à la mesure de Lebesgue sur $\mathbb{R}^d$ et de densité f_X. Alors les X_i sont des v.a. à loi absolument continue par rapport à la mesure de Lebesgue sur $\mathbb{R}$ de densité $f_{X_i}(x_i) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_X(z_1, \ldots, z_{i-1}, x_i, z_{i+1}, \ldots, z_d)\, dz_1 \ldots dz_{i-1} dz_{i+1} \ldots dz_d$.*

Proof. Si F_X est la fonction de répartition de X, alors pour tout $x_i \in \mathbb{R}$, $F_{X_i}(x_i) = \lim_{x_j \to \infty, j \neq i} F_X(x_1, \ldots, x_d)$. Mais comme la mesure est absolument continue par rapport à la mesure de Lebesgue sur $\mathbb{R}^d$, il existe f_X tel que $F_X(x_1, \ldots, x_d) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_d} f_X(z_1, \ldots, z_d)\, dz_1 \ldots dz_d$. Donc en utilisant Fubini

$$\begin{aligned} F_{X_i}(x_i) &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{x_i} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_X(z_1, \ldots, z_{i-1}, z_i, z_{i+1}, \ldots, z_d)\, dz_1 \ldots dz_d \\ &= \int_{-\infty}^{x_i} \Big(\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_X(z_1, \ldots, z_{i-1}, z_i, z_{i+1}, \ldots, z_d)\, dz_1 \ldots dz_{i-1} dz_{i+1} \ldots dz_d \Big) dz_i \\ &= \int_{-\infty}^{x_i} f_{X_i}(z_i)\, dz_i, \end{aligned}$$

avec $f_{X_i}(z_i) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_X(z_1, \ldots, z_{i-1}, z_i, z_{i+1}, \ldots, z_d)\, dz_1 \ldots dz_{i-1} dz_{i+1} \ldots dz_d$, fonction mesurable positive: la loi de X_i est bien absolument continue par rapport à la mesure de Lebesgue sur $\mathbb{R}$ et sa densité est f_{X_i}. □

Remarque: Si X est à valeurs dans $\{x_i\}_{i\in I}$, où $x_i \in \mathbb{R}^d$ et $I \subset \mathbb{N}$, alors X est discrète et sa densité par rapport à la mesure de comptage sur $\{x_i\}_{i\in I}$ est donnée par $\mathbb{P}(X = x_i) = \mathbb{P}\big(X_1 = x_{i,1} \cap \cdots, X_d = x_{i,d}\big)$.

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

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

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

Remarque: La fonction caractéristique existe sur $\mathbb{R}$ et $\phi_X(0) = 1$. ϕ_X est aussi la transformée de Fourier de la mesure $\mathbb{P}_X$.

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

Proof. Même type de preuve que dans $\mathbb{R}$ sauf que dans le cas multidimensionnel on choisit des $a_j \leq b_j$ où $j = 1, \ldots, n$ vérifiant $\mathbb{P}_X\big(\{a_1, b_1\} \times \cdots \times \{a_d, b_d\}\big) = 0$, et alors on montre que

$$\mathbb{P}(X \in [a_1, b_1] \times \cdots \times [a_d, b_d]) = \lim_{T \to \infty} \prod_{j=1}^{d} \Big(\frac{1}{2\,\pi} \int_{-T}^{T} \frac{e^{-it_j a_j} - e^{-it_j b_j}}{it_j} \Big)\, \phi_X(t_1, \ldots, t_d)\, dt_1 \ldots dt_d.$$

□

Théorème (Théorème d'inversion)**.** *Si X est un vecteur aléatoire sur $(\Omega, \mathcal{A}, \mathbb{P})$ 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 $x \in \mathbb{R}^d$,*

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

Proof. Même type de preuve que dans $\mathbb{R}$. □

Corollaire. (Lemme de Cramèr-Wold) *Si X et Y sont deux vecteurs aléatoires définis sur $(\Omega, \mathcal{A}, \mathbb{P})$ à valeurs dans $\mathbb{R}^d$. Alors $\mathbb{P}_X = \mathbb{P}_Y$ si et seulement si $< u, X > \overset{\mathcal{L}}{\sim} < u, Y >$ pour tous les $u \in \mathbb{R}^d$.*

Ceci signifie que la loi d'un vecteur aléatoire $X = (X_1, \ldots, X_d)$ peut être déterminée à partir de la loi de des combinaisons linéaires de ses coordonnées qui sont les variables aléatoires X_i. Cela peut être intéressant pour montrer notamment des théorèmes de la limite centrale multidimensionnels: on pourra se restreindre à montrer un théorème de la limite centrale unidimensionnel pour une combinaison linéaire générale.

Définition. Pour X un vecteur aléatoire sur $(\Omega, \mathcal{A}, \mathbb{P})$ à valeurs dans $\mathbb{R}^d$ telle que $\mathbb{E}\big[\|X\|^r\big] < \infty$ avec $r > 0$.

1. Si $r = 1$, alors on définit le vecteur $\mathbb{E}[X] = \big(\mathbb{E}[X_1], \ldots, \mathbb{E}[X_d]\big)^\top$, qui est appelé **espérance** de X;
2. Si $r = 2$, alors on définit la matrice $\mathrm{cov}(X) = \mathbb{E}\big[\big(X - \mathbb{E}[X]\big)\,\big(X - \mathbb{E}[X]\big)^\top\big] = \big(\mathrm{cov}(X_i, X_j)\big)_{1 \leq i,j \leq d}$ appelée **matrice de variance-covariance** de X.

Propriété. *Soit X et Y deux vecteurs aléatoires définis sur $(\Omega, \mathcal{A}, \mathbb{P})$ à valeurs dans $\mathbb{R}^d$.*

1. *Si A et B sont des matrices de réels de taille (ℓ, d) et C un vecteur de $\mathbb{R}^d$, alors:*

$$\mathbb{E}\big[A\,X + B\,Y + C\big] = A\,\mathbb{E}[X] + B\,\mathbb{E}[Y] + C.$$

2. *Si A est une matrice de réels de taille (ℓ, d) et C un vecteur de $\mathbb{R}^d$,*

$$cov\big(A\,X + C\big) = A\,cov(X)\,A^\top.$$

Proof. 1. Voir la dimension 1.

2. On a $\mathbb{E}\big[A\,X+C\big] = A\,\mathbb{E}[X]+C$ d'où:

$$\mathrm{cov}\,(A\,X+C) = \mathbb{E}\big[((A\,X+C)-\mathbb{E}[A\,X+C])\,((A\,X+C)-\mathbb{E}[A\,X+C])^{\top}\big] = \mathbb{E}\big[A\,(X-\mathbb{E}[X])\,{}^t(A\,(X-\mathbb{E}[X]))^{\top}\big] = A\,\mathrm{cov}(X)\,A^{\top}.$$

□

Propriété. *Pour X un vecteur aléatoire sur $(\Omega,\mathcal{A},\mathbb{P})$ à valeurs dans $\mathbb{R}^d$ telle que $\mathbb{E}\big[\|X\|^2\big] < \infty$ alors sa matrice de variance-covariance Σ est une matrice symétrique positive.*

Proof. Comme $\mathrm{cov}(X_i,X_j) = \mathrm{cov}(X_j,X_i)$ la matrice Σ est clairement symétrique. Elle est donc la matrice d'une forme quadratique et est diagonalisable dans $\mathbb{R}$. De plus, pour tout $u\in\mathbb{R}^d$, ${}^t\!u\,X$ est une v.a. et sa variance $\mathrm{var}({}^t\!u\,X)\geq 0$, comme toute variance. Mais $\mathrm{var}({}^t\!u\,X) = \mathrm{cov}({}^t\!u\,X) = {}^t\!u\,\mathrm{cov}(X)\,u$ par la formule précédente. Donc ${}^t\!u\,\mathrm{cov}(X)\,u\geq 0$ pour tout $u\in\mathbb{R}^d$: la matrice est positive (les valeurs propres sont donc toutes positives ou nulles). □

2.2 Indépendance

Définition. Soit $(\Omega,\mathcal{A},\mathbb{P})$ un espace de probabilité et $I\subset\mathbb{N}$.

- Soit $(A_i)_{i\in I}$ une famille 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 $K\subset I$,
$$\mathbb{P}\left(\bigcap_{i\in K} A_i\right) = \prod_{i\in K}\mathbb{P}(A_i).$$
- 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 $(X_i)_{i\in I}$ une famille de variables aléatoires sur $(\Omega,\mathcal{A},\mathbb{P})$ 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, ou encore pour tous les sous-ensembles finis $K\subset I$:
$$\mathbb{P}\big(\bigcap_{i\in K} X_i\in B_i\big) = \prod_{i\in K}\mathbb{P}(X_i\in B_i)\quad\text{pour tous } B_i\in\mathcal{B}(\mathbb{R}).$$

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

Proof. Par le théorème du transport et Fubini, il est clair que pour toute famille $(B_i)_{1\leq i\leq n}$ de boréliens de $\mathcal{B}(\mathbb{R})$ alors:

$$\mathbb{P}(X_1\in B_1,\ldots,X_n\in B_n) = \int_{B_1\times\cdots\times B_n} d\mathbb{P}_{(X_1,\ldots,X_n)}(x_1,\ldots,x_n) = \prod_{i=1}^{n}\int_{B_i} d\mathbb{P}_{X_i}(x_i) = \prod_{i=1}^{n}\mathbb{P}(X_i\in B_i).$$

□

Proposition. *Si $(X_i)_{i\in I}$ sont des variables aléatoires indépendantes sur $(\Omega,\mathcal{A},\mathbb{P})$. 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*

$$\mathbb{E}\Big[\prod_{j\in J} g_j(X_j)\Big] = \prod_{j\in J}\mathbb{E}[g_j(X_j)].$$

Proof. $\Longrightarrow$ Par le théorème du transport, puis par Fubini,

$$\mathbb{E}\Big[\prod_{j\in J} g_j(X_j)\Big] = \int_{\mathbb{R}^{|J|}}\prod_{j\in J} g_j(x_j)\,d\mathbb{P}_{(X_j)_{j\in J}}((dx_j)_{j\in J}) = \prod_{j\in J}\int_{\mathbb{R}} g_j(x_j)\,d\mathbb{P}_{X_j}(dx_j) = \prod_{j\in J}\mathbb{E}[g_j(X_j)].$$

$\Longleftarrow$ On prenant le cas particulier $g_j(x_j) = \mathbb{I}_{x_j\in B_j}$, avec (B_j) des boréliens de $\mathcal{B}(\mathbb{R})$ on retombe sur la définition de l'indépendance. □

Corollaire. *$(X_1, \cdots, X_d)$ sont des variables aléatoires indépendantes si et seulement si pour tout $(t_1, \cdots, t_d) \in \mathbb{R}^d$,*

$$\phi_{(X_1,\cdots,X_d)}(t_1, \cdots, t_d) = \prod_{j=1}^{d} \phi_{X_j}(t_j).$$

Proof. $\Longrightarrow$ Comme $\phi_{(X_1,\cdots,X_d)}(t_1, \cdots, t_d) = \mathbb{E}\big[e^{i \sum_{i=1}^d t_j X_j}\big] = \mathbb{E}\big[\prod_{j=1}^d e^{i t_j X_j}\big]$, on utilise la caractérisation précédente de l'indépendance avec les $g_j(x) = e^{i t_j x}$.
$\Longleftarrow$ D'après la formule de caractérisation de la loi par la fonction caractéristique, avec $\mathbb{P}(\{a_i, b_i\}_{1\le i\le d}) = 0$,

$$\begin{aligned}\mathbb{P}(X \in [a_1, b_1] \times \cdots \times [a_d, b_d]) &= \lim_{T\to\infty} \int_{-T}^{T} \cdots \int_{-T}^{T} \prod_{j=1}^{d} \Big(\frac{1}{2\,\pi} \int_{-T}^{T} \frac{e^{-it_j a_j} - e^{-it_j b_j}}{it_j} \Big)\, \phi_X(t_1, \ldots, t_d)\, dt_1 \ldots dt_d \\ &= \lim_{T\to\infty} \prod_{j=1}^{d} \Big(\frac{1}{2\,\pi} \int_{-T}^{T} \frac{e^{-it_j a_j} - e^{-it_j b_j}}{it_j} \phi_{X_j}(t_j) dt_j \Big) = \prod_{j=1}^{d} \mathbb{P}(X_j \in [a_j, b_j]),\end{aligned}$$

soit la caractérisation de l'indépendance pour presque tous les pavés fermés. □

Corollaire. *Soit $X = (X_1, \ldots, X_n)$ un vecteur aléatoire défini sur $(\Omega, \mathcal{A}, \mathbb{P})$ à valeurs dans $\mathbb{R}^d$ dont la mesure de probabilité $\mathbb{P}_X$ est absolument continue par rapport à la mesure de Lebesgue λ_d sur $\mathbb{R}^d$ avec pour densité f_X. Alors:*

$$(X_1, \cdots, X_n) \text{ v.a. indépendantes} \iff f_X(x_1, \ldots, x_n) = \prod_{i=1}^{d} f_{X_i}(x_i) \text{ pour tout } (x_1, \ldots, x_n) \in \mathbb{R}^n.$$

Proof. On utilise la formule d'inversion pour la fonction caractéristique:

$$f_X(x_1, \ldots, x_n) = \frac{1}{(2\,\pi)^n} \int_{\mathbb{R}^n} \phi_X(u_1, \ldots, u_n)\, e^{-i(u_1 x_1 + \cdots + u_n x_n)} du_1 \ldots du_n = \frac{1}{(2\,\pi)^n} \int_{\mathbb{R}^n} \prod_{j=1}^{n} \phi_{X_j}(u_j)\, e^{-i\,u_j x_j} du_1 \ldots du_n$$

d'après la caractérisation précédente. Il ne reste plus qu'à utiliser Fubini et à réécrire l'intégrale multiple comme un produit d'intégrales simples. □

Corollaire. *Deux vecteurs aléatoires définis sur $(\Omega, \mathcal{A}, \mathbb{P})$ sont indépendants si et seulement si toute combinaison linéaire de l'un est indépendante de toute combinaison linéaire de l'autre.*

Corollaire. *Soit $X = (X_1, \ldots, X_n)$ un vecteur aléatoire défini sur $(\Omega, \mathcal{A}, \mathbb{P})$ à valeurs dans $\mathbb{R}^d$. On suppose que $\mathbb{E}\big[\|X\|^2\big] < \infty$ et $(X_1, \ldots, X_n)$ mutuellement indépendantes. Alors:*

$$cov(X) \text{ est une matrice diagonale avec les } var(X_i) \text{ sur sa diagonale.}$$

Proof. On utilise ici le fait que $\mathrm{cov}(X) = (\mathrm{cov}(X_i, X_j))_{1\le i,j\le n}$ et $\mathrm{cov}(X_i, X_j) = 0$ si $i \neq j$. □

Remarque: Il est bien connu que $\mathrm{cov}(X_1, X_2) = 0$ n'entraîne pas que X_1 et X_2 soient indépendantes. Exemple: $X_1 \overset{\mathcal{L}}{\sim} \mathcal{N}(0, 1)$ et $X_2 = X_1^2$.

Propriété. *Soit X et Y deux v.a.* **indépendantes** *et de loi absolument continues par rapport à la mesure de Lebesgue λ_1, de densités f_X et f_Y. Alors $Z = X + Y$ est une v.a. absolument continue par rapport à la mesure de Lebesgue λ_1 de densité f_Z et*

$$f_Z(z) = \int_{\mathbb{R}} f_X(t)\, f_Y(z - t)\, dt = \int_{\mathbb{R}} f_Y(t)\, f_X(z - t)\, dt \quad \text{pour tout } z \in \mathbb{R}.$$

f_Z est appelée **produit de convolution de f_X et f_Y**.

Proof. On peut écrire en utilisant Fubini et des changements de variables, que pour tout $z \in \mathbb{R}$,

$$F_{X+Y}(z) = \mathbb{P}(X+Y \le z) = \int_{-\infty}^{\infty} f_Y(y) \int_{-\infty}^{z-y} f_X(x)\, dx\, dy = \int_{-\infty}^{\infty} \int_{-\infty}^{z} f_Y(y) f_X(x'-y)\, dx'\, dy = \int_{-\infty}^{z} \Big(\int_{-\infty}^{\infty} f_Y(y) f_X(x'-y)\, dy \Big)\, dx',$$

d'où le résultat. □

Remarque: Attention l'hypothèse d'indépendance est nécessaire pour obtenir le produit de convolution.

3 Vecteurs gaussiens

3.1 Définitions et premières propriétés

Définition. On dit qu'une variable aléatoire X est gaussienne si c'est une variable absolument continue par rapport à la mesure de Lebesgue sur $\mathbb{R}$ de densité

$$f(x) = \frac{1}{\sqrt{2\,\pi}}\,\frac{1}{\sigma}\,\exp\Big(-\frac{1}{2}\,\frac{(x-m)^2}{\sigma^2}\Big) \quad \text{avec } m \in \mathbb{R} \text{ et } \sigma^2 > 0.$$

On note $X \overset{\mathcal{L}}{\sim} \mathcal{N}(m, \sigma^2)$ et $m = \mathbb{E}[X]$ et $\sigma^2 = \mathrm{var}(X)$.

Propriété. *Si $X \overset{\mathcal{L}}{\sim} \mathcal{N}(m, \sigma^2)$, alors:*

1. *$X = m + \sigma\, Z$ avec $Z \overset{\mathcal{L}}{\sim} \mathcal{N}(0,1)$, loi normale centrée réduite.*
2. *$\mathbb{E}[|X|^p] = f_p(m, \sigma^2) < \infty$ pour tout $p \in \mathbb{N}^*$.*
3. *La fonction caractéristique de X est $\phi_X(u) = \exp\big(i\,m\,u - \frac{1}{2}\,\sigma^2\,u^2\big)$.*

Définition. On dit qu'un vecteur aléatoire $(X_1, \ldots, X_d)$ non nul défini sur $(\Omega, \mathcal{A}, \mathbb{P})$ est un vecteur gaussien si toute combinaison linéaire $u_1X_1 + \cdots + u_dX_d$, avec $(u_i)_{1\le i\le d} \in \mathbb{R}^d$, est une variable gaussienne ou une constante.

Conséquence. Si $(X_1, \ldots, X_d)$ est un vecteur gaussien, alors chaque X_i est une variable gaussienne. La réciproque est fausse en général (voir exercice).

Conséquence. Si $X = {}^t(X_1, \ldots, X_d)$ est un vecteur gaussien, alors pour toute matrice M de réels de taille (d', d) et tout vecteur colonne B de taille d', $M\,X + B$ est aussi un vecteur gaussien.

Propriété. *Si $X = (X_1, \ldots, X_d)$ est un vecteur gaussien, alors sa loi de probabilité ne dépend que de son espérance et de sa matrice de variance-covariance.*

Proof. Commençons par étudier le cas où X est centré. On a $\phi_X(u) = \mathbb{E}\big[e^{i\,\sum_{j=1}^d u_j\,X_j}\big]$. Comme $\sum_{j=1}^d u_j\,X_j$ est une variable aléatoire gaussienne centrée, sa fonction caractéristique dépend de sa variance et $\mathrm{var}\big(\sum_{j=1}^d u_j\,X_j\big) = \sum_{1\le j,k\le d} u_j u_k\,\mathrm{cov}(X_j, X_k)$. On en déduit que:

$$\phi_X(u) = \exp\Big(-\frac{1}{2}\,\big(\sum_{1\le j,k\le d} u_j u_k\,\mathrm{cov}(X_j, X_k)\big)\Big) = \exp\Big(-\frac{1}{2}\,{}^t u\,\Sigma\,u\Big),$$

où $\Sigma = \big(\mathrm{cov}(X_j, X_k)\big)_{1\le j,k\le d}$ est la matrice de covariance de X.
Si maintenant, X n'est plus centré, si on note $m = {}^t(m_1, \ldots, m_d)$ son espérance, on peut écrire que

$$\phi_X(u) = \mathbb{E}\big[e^{i<u,m>}\,e^{i\,\sum_{j=1}^d u_j\,(X_j - m_j)}\big] = e^{i<u,m>}\,\exp\Big(-\frac{1}{2}\,{}^t u\,\Sigma\,u\Big).$$

La fonction caractéristique de X ne dépend donc que de son espérance et sa matrice de covariance, et il en est de même pour sa loi de probabilité. □

Conséquence: Si $X \overset{\mathcal{L}}{\sim} \mathcal{N}\big(m\,,\,\Sigma\big)$ alors $X = m + \Sigma^{1/2}\,Z$, avec $Z \overset{\mathcal{L}}{\sim} \mathcal{N}\big(0_d\,,\,I_d\big)$ vecteur gaussien centré réduit (la matrice $\Sigma^{1/2}$ peut être obtenue par diagonalisation avec racines des valeurs propres).

Propriété. *Si $X_1, \ldots, X_d$ sont d v.a. gaussiennes* **indépendantes** *alors $X = (X_1, \ldots, X_d)$ est un vecteur gaussien de matrice de variance-covariance diagonale.*

Proof. On a vu que l'indépendance entre 2 v.a. entraîne la nullité de leur covariance. Or la matrice de covariance est composée en dehors de la diagonale des $\mathrm{cov}(X_j, X_k)$ avec $k \neq j$. Si on considère la fonction caractéristique de X, alors, en utilisant l'indépendance des X_j:

$$\phi_X(u) = \prod_{j=1}^{d} \phi_{X_j}(u_j) = e^{i<u,m>}\, e^{-\frac{1}{2}\sum_{j=1}^{d}\sigma_j^2 u_j^2} = e^{i<u,m>}\, \exp\Big(-\frac{1}{2}\,{}^t u\,\Sigma\, u\Big),$$

avec Σ la matrice de covariance de X qui est donc diagonale: on aboutit à la fonction caractéristique d'un vecteur gaussien. □

Corollaire. *Soit $X = (X_1, \ldots, X_d)$ un vecteur gaussien. Alors:*

$$(X_1, \ldots, X_d) \textit{ indépendantes } \iff \textit{ cov}(X_i, X_j) = 0 \textit{ pour } i \neq j.$$

Proof. Dans le sens $\Longrightarrow$ c'est immédiat. Dans le sens $\Longleftarrow$ on reprend la preuve un peu plus haut. □

Propriété. *Si $X = (X_1, \ldots, X_d)$ est un vecteur gaussien d'espérance $\mathbb{E}[X]$ dont la matrice de variance-covariance Σ est définie positive, alors X est absolument continue par rapport à la mesure de Lebesgue sur $\mathbb{R}^d$ et sa densité est:*

$$f_X(x) = \frac{1}{(2\pi)^{d/2}}\,\frac{1}{\sqrt{\det(\Sigma)}}\,\exp\Big(-\frac{1}{2}\,\big(x-\mathbb{E}[X]\big)^\top \Sigma^{-1}\big(x-\mathbb{E}[X]\big)\Big) \quad \textit{pour } x = (x_1, \ldots, x_d)^\top \in \mathbb{R}^d.$$

Proof. En premier lieu, on obtient de manière immédiate que si le vecteur Z est centré réduit, soit $Z \overset{\mathcal{L}}{\sim} \mathcal{N}(0, I_d)$ alors toutes ses composantes sont indépendantes donc sa densité est $\displaystyle\prod_{j=1}^{d} \frac{1}{(2\pi)^{1/2}}\, e^{-\frac{1}{2}z_j^2} = \frac{1}{(2\pi)^{d/2}}\, e^{-\frac{1}{2}\sum_{j=1}^{d} z_j^2}$.

Par ailleurs, on a vu que si $X \overset{\mathcal{L}}{\sim} \mathcal{N}(\mathbb{E}[X], \Sigma)$ alors $X \overset{\mathcal{L}}{\sim} \mathbb{E}[X] + \Sigma^{1/2} Z$. Ainsi pour toute fonction $g : \mathbb{R}^d \to \mathbb{R}$ bornée, alors:

$$\mathbb{E}[g(X)] = \mathbb{E}\big[g(\mathbb{E}[X] + \Sigma^{1/2} Z)\big] = \int_{\mathbb{R}^d} g(\mathbb{E}[X] + \Sigma^{1/2} z)\, f_Z(z)\, dz = \int_{\mathbb{R}^d} g(x)\, \det\big(\Sigma^{-1/2}\big)\, f_Z\big(\Sigma^{-1/2}(x - \mathbb{E}[X])\big)\, dx$$

avec la formule du changement de variable pour les intégrales multiples. Comme $\det\big(\Sigma^{-1/2}\big) = \big(\det(\Sigma)\big)^{-1/2}$ et $\big(\Sigma^{-1/2}(x - \mathbb{E}[X])\big)^\top \Sigma^{-1/2}(x - \mathbb{E}[X]) = (x - \mathbb{E}[X]))^\top \Sigma^{-1}(x - \mathbb{E}[X])$ on obtient le résultat. □

3.2 Lois découlant de la loi gaussienne

Définition. Soit $X = (X_1, \ldots, X_d)$ un vecteur gaussien centré réduit, $X \overset{\mathcal{L}}{\sim} \mathcal{N}(0, I_d)$. Alors la loi de $Z = \|X\|^2 = X_1^2 + \cdots + X_d^2$ est appelée **loi du chi2 à d degrés de liberté** notée $Z \overset{\mathcal{L}}{\sim} \chi^2(d)$.

Propriété. *Pour $d \in \mathbb{N}^*$, $\chi^2(d) \overset{\mathcal{L}}{\sim} \Gamma\big(\frac{d}{2}, \frac{1}{2}\big)$.*

Proof. On sait que si $U \overset{\mathcal{L}}{\sim} \Gamma(\alpha, \beta)$ et $v \overset{\mathcal{L}}{\sim} \Gamma(\alpha, \beta)$ et si $U \perp\!\!\!\perp V$, alors $U + V \overset{\mathcal{L}}{\sim} \Gamma(2\,\alpha, \beta)$. Il suffit donc de démontrer que $X_1^2 \overset{\mathcal{L}}{\sim} \Gamma\big(\frac{1}{2}, \frac{1}{2}\big)$.
Or pour $x \geq 0$,

$$\mathbb{P}(X_1^2 \leq x) = 2\,\mathbb{P}(0 \leq X_1 \leq \sqrt{x}) = 2\,\big(F(\sqrt{x}) - \frac{1}{2}\big) \implies f_{X_1^2}(x) = \frac{1}{\sqrt{2\pi}\,\sqrt{x}}\, e^{-x/2} = \frac{\big(\frac{1}{2}\big)^{1/2}}{\Gamma(1/2)}\, x^{\frac{1}{2}-1}\, e^{-\frac{1}{2}x},$$

qui est bien la loi $\Gamma\big(\frac{1}{2}, \frac{1}{2}\big)$. □

Théorème (Théorème de Cochran). *Soit un vecteur gaussien centré réduit $X \overset{\mathcal{L}}{\sim} \mathcal{N}_d(0_d\,, I_d)$. En considérant le produit scalaire euclidien classique sur $\mathbb{R}^d$, on considère A et B deux sous-espaces vectoriels orthogonaux de $\mathbb{R}^d$ et on note P_A et P_B les matrices de projection orthogonale sur A et sur B. Alors:*

1. *$P_A X$ et $P_B X$ sont deux vecteurs gaussiens* **indépendants** *à valeurs dans $\mathbb{R}^d$;*
2. *$\|P_A X\|^2 \overset{\mathcal{L}}{\sim} \chi^2(\dim(A))$, loi du χ^2 à $\dim(A)$ degrés de liberté.*

Proof. 1. Comme P_A est une matrice de réels et ε un vecteur gaussien, alors $P_A\varepsilon$ est un vecteur gaussien, centré car ε est centré. Donc $(P_A\varepsilon, P_B\varepsilon)$ est une famille indépendante si $\mathrm{cov}(P_A\varepsilon\,,\, P_B\varepsilon) = 0$ car $(P_A\varepsilon, P_B\varepsilon)$ est un vecteur gaussien (de taille $2d$). Or $\mathrm{cov}(P_A\varepsilon\,,\, P_B\varepsilon) = \mathbb{E}[P_A\varepsilon\,{}^t(P_B\varepsilon)] = \mathbb{E}[P_A\varepsilon\,{}^t\varepsilon\,{}^tP_B] = P_A\mathrm{cov}(\varepsilon)^tP_{AB} = P_AP_B$. Mais $A \perp B$ donc $P_AP_B = 0$, d'où le résultat.

2. $\|P_A\varepsilon\|^2 = {}^t\varepsilon\,{}^tP_AP_A\varepsilon = {}^t\varepsilon\,P_A\,\varepsilon$. Mais P_A est une matrice réelle symétrique donc diagonalisable et on peut écrire que $P_A = Q\,D\,{}^tQ$ avec Q une matrice orthogonale et D une matrice diagonale avec les valeurs propres de P_A, donc par exemple d'abord $\dim(A)$ uns sur la diagonale puis dessous $n - \dim(A)$ zéros. D'où $\|P_A\varepsilon\|^2 = {}^t(Q\,\varepsilon)\,D\,(Q\,\varepsilon)$. Soit $\varepsilon' = {}^tQ\,\varepsilon$. Alors $\varepsilon' \overset{\mathcal{L}}{\sim} \mathcal{N}_n(0\,,\,{}^tQ\,Q) = \mathcal{N}_n(0\,,\,I_d)$ car Q est une matrice orthogonale (donc $Q^{-1} = {}^tQ$). Donc ε' est un vecteur gaussien centré standard. Comme ${}^t\varepsilon'\,D\,\varepsilon' = \sum_{j=1}^{\dim(A)}(\varepsilon'_j)^2$, les ε'_j étant des variables gaussiennes centrées indépendantes de même variance 1, on a donc $\sum_{j=1}^{\dim(A)}(\varepsilon'_j)^2 \overset{\mathcal{L}}{\sim} \chi^2(\dim(A))$.

□

Définition. Soit X une v.a. gaussienne de loi $\mathcal{N}(0,1)$ et Y une v.a. **indépendendante** de X et suivant une loi $\chi^2(d)$ avec $d \in \mathbb{N}^*$. Alors la variable $T = X/\sqrt{Y/d}$ suit une loi dite de Student à d degrés de liberté, et on notera $T \overset{\mathcal{L}}{\sim} t(d)$.

Propriété. *Si T suit une loi de Student à d degrés de liberté:*

- *la loi de T est symétrique et de densité* $f_T(t) = \dfrac{1}{\sqrt{d\,\pi}}\dfrac{\Gamma(\frac{d+1}{2})}{\Gamma(\frac{d}{2})}\left(1 + \dfrac{t^2}{d}\right)^{-\frac{d+1}{2}}$.
- *$\mathbb{E}[|T|^\alpha] < \infty$ pour $\alpha < d$ et $\mathbb{E}[|T|^d] = \infty$.*
- *Si $d = 1$, T suit une loi de Cauchy.*

Proof.

- Pour $t \in \mathbb{R}$, du fait que X et Y sont indépendantes et du fait que $Y \geq 0$, on peut écrire que:

$$\mathbb{P}(T \leq t) = \mathbb{P}(X \leq t\sqrt{Y/d}) = \int_0^\infty \frac{(\frac{1}{2})^{d/2}}{\Gamma(\frac{d}{2})} y^{\frac{d}{2}-1}e^{-y/2}\int_{-\infty}^{t\sqrt{y/d}} \frac{1}{\sqrt{2\pi}}e^{-x^2/2}\,dx\,dy = \frac{(\frac{1}{2})^{d/2}}{\sqrt{2\pi}\sqrt{d}\,\Gamma(\frac{d}{2})}\int_0^\infty\int_{-\infty}^{t} y^{\frac{d}{2}-\frac{1}{2}}e^{-y/2}e^{-\frac{y\,u^2}{2d}}\,du\,dy$$

  après un changement de variable. De ceci, en appliquant Fubini et un nouveau changement de variable, on en déduit que:

$$\mathbb{P}(T \leq t) = \frac{(\frac{1}{2})^{d/2}}{\sqrt{2\pi}\sqrt{d}\,\Gamma(\frac{d}{2})}\int_{-\infty}^{t}\int_0^\infty y^{\frac{(d+1)}{2}-1}e^{-\frac{1}{2}\,y\left(1+\frac{u^2}{d}\right)}dy\,du = \frac{(\frac{1}{2})^{d/2}}{\sqrt{2\pi}\sqrt{d}\,\Gamma(\frac{d}{2})}\int_{-\infty}^{t}(1+\frac{u^2}{d})^{-\frac{(d+1)}{2}}\int_0^\infty y^{\frac{(d+1)}{2}-1}e^{-\frac{1}{2}\,y}dy\,du.$$

  Il ne reste plus alors qu'à utiliser le fait que $\int_0^\infty y^{\frac{(d+1)}{2}-1}e^{-\frac{1}{2}\,y}dy = \Gamma(\frac{d+1}{2})\,(\frac{1}{2})^{-(d+1)/2}$, et on btient le résultat désiré.

- En utilisant un équivalent, on a $|t|^\alpha f_T(t) \sim |t|^\alpha\,t^{-(d+1)}$ pour $t \to \infty$ et il ne reste plus qu'à utiliser les résultats connus sur les intégrales de Riemann.
- Si $d = 1$, on tombe bien sur la densité $\frac{1}{\pi}\frac{1}{1+x^2}$ qui est celle d'une loi de Cauchy.

□

Propriété. *Soit $(X_1, \ldots, X_n)$ une suite de v.a.i.i.d. gaussienne de loi $\mathcal{N}(m, \sigma^2)$. Soit la moyenne et la variance empiriques de $(X_1, \ldots, X_n)$ définies respectivement par:*

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i \quad et \quad \overline{\sigma}_n^2 = \frac{1}{n-1}\sum_{i=1}^n \left(X_i - \overline{X}_n\right)^2.$$

Alors:

- *La loi de $\overline{X}_n$ est $\mathcal{N}\left(m\,,\,\dfrac{\sigma^2}{n}\right)$ et la loi de $\overline{\sigma}_n^2$ est $\dfrac{\sigma^2}{n-1}\,\chi^2(n-1)$.*

- *$\overline{X}_n$ et $\overline{\sigma}_n^2$ sont deux v.a. indépendantes et la loi de $\widehat{T}_n = \dfrac{\sqrt{n}(\overline{X}_n - m)}{\sqrt{\overline{\sigma}_n^2}}$ est une loi $t(n-1)$.*

Proof. Voir exercice de la feuille de TD3. □

Définition. Soit X_1 et X_2 deux v.a. **indépendantes** suivant respectivement des lois $\chi^2(n_1)$ et $\chi^2(n_2)$. Alors la variable $Z = \dfrac{\frac{1}{n_1} X_1}{\frac{1}{n_2} X_2}$ suit une loi dite de Fisher à (n_1, n_2) degrés de liberté, et on notera $Z \overset{\mathcal{L}}{\sim} \mathcal{F}(n_1, n_2)$.

4 Convergence de suites de variables aléatoires et théorèmes limite

4.1 Convergence de suites de variables aléatoires

Remarque importante! Toutes (ou presque) les définitions et propriétés données ci-dessous le sont pour des suites de variables aléatoires. Cependant elles sont également valables pour des suites de vecteurs aléatoires, quite à adapter les définitions et preuves avec $\mathbb{R}^d$ et $\|\cdot\|$ en lieu et place de $\mathbb{R}$ et $|\cdot|$.

Définition. Soit $(X_n)_{n\in\mathbb{N}}$ une suite de v.a. définie sur le même espace de probabilité $(\Omega, \mathcal{A}, \mathbb{P})$. Et soit X_∞ une v.a. définie également sur $(\Omega, \mathcal{A}, \mathbb{P})$. Alors:

- (X_n) converge en loi vers X_∞, noté $X_n \xrightarrow[n\to\infty]{\mathcal{L}} X_\infty$, lorsque,

$$\lim_{n\to\infty} F_{X_n}(x) = F_{X_\infty}(x) \quad \text{où } x \in \mathbb{R} \text{ et } F_{X_\infty} \text{ continue en } x.$$

- (X_n) converge en probabilité vers X_∞, noté $X_n \xrightarrow[n\to\infty]{\mathcal{P}} X_\infty$, lorsque pour tout $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\big(|X_n - X_\infty| \geq \varepsilon\big) = 0.$$

- (X_n) converge dans $\mathbb{L}^p(\Omega, \mathcal{A}, \mathbb{P})$ vers X_∞, noté $X_n \xrightarrow[n\to\infty]{\mathbb{L}^p} X_\infty$, avec $p > 0$, lorsque

$$\mathbb{E}\big[|X_n|^p + |X_\infty|^p\big] < \infty \quad \text{et} \quad \lim_{n\to\infty} \mathbb{E}\big[|X_n - X_\infty|^p\big] = 0.$$

- (X_n) converge presque sûrement vers X_∞, noté $X_n \xrightarrow[n\to\infty]{p.s.} X_\infty$, lorsque

$$\lim_{n\to\infty} X_n(\omega) = X_\infty(\omega) \quad \text{pour } \mathbb{P}\text{-presque tout } \omega \in \Omega.$$

Théorème. *Soit $(X_n)_{n\in\mathbb{N}}$ est une suite de v.a. définies sur $(\Omega, \mathcal{A}, \mathbb{P})$ et X_∞ une v.a. également définie sur $(\Omega, \mathcal{A}, \mathbb{P})$. Alors,*

$$X_n \xrightarrow[n\to\infty]{\mathcal{L}} X_\infty \iff \mathbb{E}[g(X_n)] \underset{n\to\infty}{\longrightarrow} \mathbb{E}[g(X_\infty)] \text{ pour toute fonction } g : \mathbb{R} \to \mathbb{R} \text{ continue bornée.}$$

Proof. $\Longrightarrow$ Soit g une fonction continue bornée sur $\mathbb{R}$ et sans perte de généralité on va supposer que $\|g\|_\infty \leq 1$. Soit $\varepsilon > 0$. Comme $\lim_{x\to+\infty} \mathbb{P}(X > x) = 0 = \lim_{x\to-\infty} \mathbb{P}(X < x)$, il existe un intervalle $[a, b]$ de $\mathbb{R}$ tel que $\mathbb{P}(X_\infty \notin [a, b]) \leq \varepsilon$, avec F_{X_∞} continue en a et b. Sur $[a, b]$, qui est un compact, g est uniformément continue, ce qui signifie qu'il existe $\eta_\varepsilon > 0$ tel que pour tout $x_0 \in [a, b]$ et tout $x \in [x_0 - \eta_\varepsilon, x_0 + \eta_\varepsilon]$ alors $|g(x_0) - g(x)| \leq \varepsilon$]. On peut donc décomposer $[a, b]$ en $m_\varepsilon = |b - a|/\eta_\varepsilon$ intervalles $[a_i, b_i]$, soit $[a, b] = \bigcup_{i=1}^{m_\varepsilon} [a_i, b_i]$, avec pour tout $(x, y) \in [a_i, b_i]^2$, $|g(x) - g(y)| \leq \varepsilon$ et F_{X_∞} continue en a_i et b_i.
Soit g_ε la fonction en escalier telle que $g_\varepsilon(x) = \sum_{i=1}^{m_\varepsilon} g(x_i)\, \mathbb{I}_{x\in[a_i,b_i]}$ pour tout $x \in \mathbb{R}$, où $x_i \in [a_i, b_i]$ pour tout $i = 1, \ldots, m_\varepsilon$. Il est clair que pour tout $x \in [a, b]$, $|g(x) - g_\varepsilon(x)| \leq \varepsilon$. Ainsi:

$$\big|\mathbb{E}\big[g(X_n) - g_\varepsilon(X_n)\big]\big| \leq \mathbb{E}\big[\big|g(X_n) - g_\varepsilon(X_n)\big|\, \mathbb{I}_{X_n\in[a,b]}\big] + \mathbb{E}\big[\big|g(X_n) - g_\varepsilon(X_n)\big|\, \mathbb{I}_{X_n\notin[a,b]}\big] \leq \varepsilon + \mathbb{P}(X_n \notin [a, b])$$

car $g_\varepsilon = 0$ en dehors de $[a,b]$ et $\|g\|_\infty \leq 1$. De la même manière,

$$\left|\mathbb{E}\left[g(X_\infty) - g_\varepsilon(X_\infty)\right]\right| \leq \varepsilon + \mathbb{P}(X_\infty \notin [a,b])$$

Il est clair que comme $F_{X_n} \underset{n\to\infty}{\longrightarrow} F_X$, alors $\mathbb{P}(X_n \notin [a,b]) \underset{n\to\infty}{\longrightarrow} \mathbb{P}(X_\infty \notin [a,b])$, donc il existe N tel que pour $n \geq N$, $\mathbb{P}(X_n \notin [a,b]) \leq 2\varepsilon$.
Enfin, il existe N' tel que pour tout $n \geq N'$,

$$\left|\mathbb{E}\left[g_\varepsilon(X_n) - g_\varepsilon(X_\infty)\right]\right| \leq \sum_{i=1}^{m_\varepsilon} |g(x_i)| \left|\mathbb{P}(X_n \in [a_i,b_i]) - \mathbb{P}(X_\infty \in [a_i,b_i])\right| \leq \varepsilon,$$

car il existe N' tel que pour tout i, $\left|\mathbb{P}(X_n \in [a_i,b_i]) - \mathbb{P}(X_\infty \in [a_i,b_i])\right| \leq \frac{1}{m_\varepsilon}\varepsilon$. Il ne reste plus qu'à écrire que pour $n \geq \max(N,N')$,

$$\left|\mathbb{E}[g(X_n)] - \mathbb{E}[g(X_\infty)]\right| \leq \left|\mathbb{E}[g(X_n) - g_\varepsilon(X_n)]\right| + \left|\mathbb{E}[g_\varepsilon(X_\infty) - g_\varepsilon(X_n)]\right| + \left|\mathbb{E}[g(X_\infty) - g_\varepsilon(X_\infty)]\right| \leq 3\,\varepsilon + \varepsilon + 2\,\varepsilon \leq 6\,\varepsilon,$$

d'où la convergence demandée.

□

Théorème (Théorème de Lévy (1920)). *Soit $(X_n)_{n\in\mathbb{N}}$ est une suite de v.a. définies sur $(\Omega,\mathcal{A},\mathbb{P})$ et X_∞ une v.a. également définie sur $(\Omega,\mathcal{A},\mathbb{P})$. Alors, en notant ϕ_{X_n} et ϕ_{X_∞} les fonctions caractéristiques de X_n et de X_∞,*

$$X_n \underset{n\to\infty}{\xrightarrow{\mathcal{L}}} X_\infty \iff \phi_{X_n}(u) \underset{n\to\infty}{\longrightarrow} \phi_{X_\infty}(u) \text{ pour tout } u \in \mathbb{R}.$$

Proof. $\Longrightarrow$ Il suffit de considérer le théorème précédent et choisir $g(x) = e^{iux}$ qui est bien continue et bornée.
$\Longleftarrow$ Trop difficile. □

Propriété. *Soit $(X_n)_{n\in\mathbb{N}}$ une suite de v.a. définie sur le même espace de probabilité $(\Omega,\mathcal{A},\mathbb{P})$ et X_∞ une v.a. définie également sur $(\Omega,\mathcal{A},\mathbb{P})$. Alors:*

1. $X_n \underset{n\to\infty}{\xrightarrow{p.s.}} X_\infty$ *ou* $X_n \underset{n\to\infty}{\xrightarrow{\mathbb{L}^p}} X_\infty \Longrightarrow X_n \underset{n\to\infty}{\xrightarrow{\mathcal{P}}} X_\infty$.
2. $X_n \underset{n\to\infty}{\xrightarrow{\mathcal{P}}} X_\infty \Longrightarrow X_n \underset{n\to\infty}{\xrightarrow{\mathcal{L}}} X_\infty$.
3. *pour* $q \geq p$, $X_n \underset{n\to\infty}{\xrightarrow{\mathbb{L}^p}} X_\infty \Longrightarrow X_n \underset{n\to\infty}{\xrightarrow{\mathbb{L}^q}} X_\infty$.
4. *La convergence en loi n'entraîne pas la convergence en probabilité. Mais pour C une constante,* $(X_n \underset{n\to\infty}{\xrightarrow{\mathcal{P}}} C) \Longleftrightarrow (X_n \underset{n\to\infty}{\xrightarrow{\mathcal{L}}} C)$.

Proof. 1. • Si $X_n \underset{n\to\infty}{\xrightarrow{p.s.}} X_\infty$ alors $\exists\widetilde{\Omega} \in \mathcal{A}$ vérifiant $\mathbb{P}(\widetilde{\Omega}) = 1$, tel que pour tout $\omega \in \widetilde{\Omega}$, $X_n(\omega) \underset{n\to\infty}{\longrightarrow} X_\infty(\omega)$. Soit $\varepsilon > 0$. Alors par la formule des probabilités totales

$$\begin{aligned}\mathbb{P}(|X_n - X_\infty| \geq \varepsilon) &= \mathbb{P}(\{\omega \in \widetilde{\Omega}, |X_n(\omega) - X_\infty(\omega)| \geq \varepsilon\}) + \mathbb{P}(\{\omega \notin \widetilde{\Omega}, |X_n(\omega) - X_\infty(\omega)| \geq \varepsilon\}) \\ &\leq \mathbb{P}(\{\omega \in \widetilde{\Omega}, |X_n(\omega) - X_\infty(\omega)| \geq \varepsilon\}) + 1 - \mathbb{P}(\widetilde{\Omega}) \leq \mathbb{P}(\{\omega \in \widetilde{\Omega}, |X_n(\omega) - X_\infty(\omega)| \geq \varepsilon\}).\end{aligned}$$

Mais comme $X_n(\omega) \underset{n\to\infty}{\longrightarrow} X_\infty(\omega)$ pour $\omega \in \widetilde{\Theta}$, alors il existe $N \in \mathbb{N}$, tel que pour tout $n \geq N$, $|X_n(\omega) - X_\infty(\omega)| < \varepsilon$. Donc pour $n \geq N$, $\{\omega \in \widetilde{\Omega}, |X_n(\omega) - X_\infty(\omega)| \geq \varepsilon\} = \emptyset$ soit $\mathbb{P}(|X_n - X_\infty| \geq \varepsilon) = 0$: d'où la convergence en probabilité.
• $X_n \underset{n\to\infty}{\xrightarrow{\mathbb{L}^p}} X_\infty$, alors $\mathbb{E}\left[|X_n|^p + |X_\infty|^p\right] < \infty$ et $\mathbb{E}[|X_n - X_\infty|^p] \underset{n\to\infty}{\longrightarrow} 0$. Pour $\varepsilon > 0$,

$$\mathbb{P}(|X_n - X_\infty| \geq \varepsilon) = \mathbb{P}(|X_n - X_\infty|^p \geq \varepsilon^p) \leq \frac{\mathbb{E}\left[|X_n - X_\infty|^p\right]}{\varepsilon^p} \underset{n\to\infty}{\longrightarrow} 0,$$

en utilisant l'Inégalité de Markov, d'où la convergence en probabilités.

2. On suppose $X_n \xrightarrow[n\to\infty]{\mathcal{P}} X_\infty$ donc pour $\eta > 0$ quelconque, $\mathbb{P}(|X_n - X_\infty| \geq \eta) \underset{n\to\infty}{\longrightarrow} 0$. Par ailleurs, pour $x \in \mathbb{R}$ tel que F_{X_∞} est continue en x, donc pour tout $\varepsilon > 0$ alors il existe $\eta_\varepsilon > 0$ tel que $|F_{X_\infty}(x - \eta_\varepsilon) - F_{X_\infty}(x)| \leq \varepsilon$ et $|F_{X_\infty}(x + \eta_\varepsilon) - F_{X_\infty}(x)| \leq \varepsilon$. Mais:
$$\begin{aligned} F_{X_n}(x) &= \mathbb{P}(X_n \leq x) \\ &= \mathbb{P}(X_\infty \leq x - (X_n - X_\infty) \cap |X_n - X_\infty| < \eta_\varepsilon) + \mathbb{P}(X_\infty \leq x - (X_n - X_\infty) \cap |X_n - X_\infty| \geq \eta_\varepsilon) \\ &\leq \mathbb{P}(X_\infty \leq x + \eta_\varepsilon) + \mathbb{P}(|X_n - X_\infty| \geq \eta_\varepsilon). \end{aligned}$$
En minorant de la même manière, on obtient donc:
$$F_{X_\infty}(x) - \varepsilon - \mathbb{P}(|X_n - X_\infty| \geq \eta_\varepsilon) \leq F_{X_n}(x) \leq F_{X_\infty}(x) + \varepsilon + \mathbb{P}(|X_n - X_\infty| \geq \eta_\varepsilon).$$
Comme $\mathbb{P}(|X_n - X_\infty| \geq \eta_\varepsilon) \underset{n\to\infty}{\longrightarrow} 0$, on en déduit qu'il existe $N \in \mathbb{N}$ tel que pour tout $n \geq N$, $\mathbb{P}(|X_n - X_\infty| \geq \eta_\varepsilon) \leq \varepsilon$, donc pour $n \geq N$, $|F_{X_n}(x) - F_{X_\infty}(x)| \leq 2\,\varepsilon$: on obtient bien la convergence en loi.
3. On utilise $\|X_n - X_\infty\|_p \leq \|X_n - X_\infty\|_q$, donc $\mathbb{E}[|X_n - X_\infty|^p] \leq \mathbb{E}[|X_n - X_\infty|^q]^{p/q}$.
4. Voir exercice.

□

Exemple instructif: Sur $([0,1], \mathcal{B}([0,1]), \mathcal{U}([0,1])$ on définit la suite (X_n) telle que pour $\omega \in [0,1]$, $X_n(\omega) = (n+1)\,\omega^n$. On montre alors que $X_n \xrightarrow[n\to\infty]{\mathcal{L}} 0$, $X_n \xrightarrow[n\to\infty]{\mathcal{P}} 0$ et $X_n \xrightarrow[n\to\infty]{p.s.} 0$, mais (X_n) ne converge pas dans $\mathbb{L}^1$ (si elle convergeait en ce sens elle convergerait également en probabilité et ce serait nécessairement vers 0).

Propriété. *Soit $(X_n)_{n\in\mathbb{N}}$ une suite de v.a. définie sur le même espace de probabilité $(\Omega, \mathcal{A}, \mathbb{P})$ et X_∞ une v.a. définie également sur $(\Omega, \mathcal{A}, \mathbb{P})$. Soit $g : \mathbb{R} \to \mathbb{R}$ est une fonction borélienne continue sur $C \subset \mathbb{R}$ telle que $\mathbb{P}(X_\infty \in C) = 1$. Alors $(X_n \xrightarrow[n\to\infty]{\mathcal{P}} X_\infty) \Longrightarrow (g(X_n) \xrightarrow[n\to\infty]{\mathcal{P}} g(X_\infty))$. Même chose pour les convergences en loi et p.s. .*

Proof. • On fixe $\varepsilon > 0$. Pour tout $\delta > 0$, on définit
$$B_{\delta,\varepsilon} = \big\{x \in \mathbb{R} \cap C,\ \exists y \in \mathbb{R},\ |x - y| \leq \delta \text{ et } |g(x) - g(y)| > \varepsilon\big\}.$$
Comme g est continue sur C, il est clair que $\lim_{\delta\to 0^+} B_{\delta,\varepsilon} = \emptyset$. En utilisant la formule des probabilités totales, on peut écrire:
$$\begin{aligned} \mathbb{P}(|g(X_n) - g(X_\infty)| > \varepsilon) &= \mathbb{P}(|g(X_n) - g(X_\infty)| > \varepsilon \cap X_\infty \in B_{\delta,\varepsilon}) + \mathbb{P}(|g(X_n) - g(X_\infty)| > \varepsilon \cap X_\infty \notin B_{\delta,\varepsilon}) \\ &\leq \mathbb{P}(X_\infty \in B_{\delta,\varepsilon}) + \mathbb{P}(|X_n - X_\infty| > \delta \cap X_\infty \notin B_{\delta,\varepsilon}) + \mathbb{P}(X_\infty \notin C) \leq \mathbb{P}(X_\infty \in B_{\delta,\varepsilon}) + \mathbb{P}(|X_n - X_\infty| > \delta), \end{aligned}$$
car $\mathbb{P}(X_\infty \notin C) = 0$. Comme $\lim_{\delta\to 0^+} B_{\delta,\varepsilon} = \emptyset$ on peut toujours choisir δ suffisamment petit pour que $\mathbb{P}(X_\infty \in B_{\delta,\varepsilon})$ soit aussi petit que l'on veut, et comme $X_n \xrightarrow[n\to\infty]{\mathcal{P}} X_\infty$, on peut choisir n suffisamment grand pour que $\mathbb{P}(|X_n - X_\infty| > \delta)$ soit aussi petit que l'on veut. Donc $g(X_n) \xrightarrow[n\to\infty]{\mathcal{P}} g(X_\infty)$.

• $X_n \xrightarrow[n\to\infty]{\mathcal{L}} X_\infty$ si et seulement si pour toute fonction h continue bornée $\mathbb{E}[h(X_n)] \underset{n\to\infty}{\longrightarrow} \mathbb{E}[h(X_\infty)]$. Comme $h_o g$ est aussi continue bornée, on a également $\mathbb{E}[h(g(X_n))] \underset{n\to\infty}{\longrightarrow} \mathbb{E}[h(g(X_\infty))]$.

• Si $X_n \xrightarrow[n\to\infty]{p.s.} X_\infty$, alors $\exists \widetilde{\Omega} \in \mathcal{A}$ vérifiant $\mathbb{P}(\widetilde{\Omega}) = 1$, tel que pour tout $\omega \in \widetilde{\Omega}$, $X_n(\omega) \underset{n\to\infty}{\longrightarrow} X_\infty(\omega)$. Donc pour tout $\omega \in \widetilde{\Omega}$, $g(X_n(\omega)) \underset{n\to\infty}{\longrightarrow} g(X_\infty(\omega))$ d'après la caractérisation de la continuité par les suites numériques. D'où $g(X_n) \xrightarrow[n\to\infty]{p.s.} g(X_\infty)$. □

Lemme (Lemme de Slutsky (Slutsky, 1915)). *Soit (X_n) une suite de v.a. définies sur $(\Omega, \mathcal{A}, \mathbb{P})$ et X_∞ une v.a. définie également sur $(\Omega, \mathcal{A}, \mathbb{P})$ telle que $X_n \xrightarrow[n\to\infty]{\mathcal{L}} X_\infty$. Soit (Y_n) une suite de v.a. définies sur $(\Omega, \mathcal{A}, \mathbb{P})$ telle que $Y_n \xrightarrow[n\to\infty]{\mathcal{P}} c$, où $c \in \mathbb{R}$. Alors $(X_n, Y_n) \xrightarrow[n\to\infty]{\mathcal{L}} (X_\infty, c)$.*

Proof. Soit $u = (u_1, u_2) \in \mathbb{R}^2$. En notant $\phi_{(X_n,Y_n)}$ la fonction caractéristique de (X_n, Y_n), on a:
$$\begin{aligned} \big|\phi_{(X_n,Y_n)}(u) - \phi_{(X_\infty,c)}(u)\big| &\leq \big|\phi_{(X_n,Y_n)}(u) - \phi_{(X_n,c)}(u)\big| + \big|\phi_{(X_n,c)}(u) - \phi_{(X_\infty,c)}(u)\big| \\ &\leq \big|\mathbb{E}\big[e^{iu_1X_n}\big(e^{iu_2Y_n} - e^{iu_2c}\big)\big]\big| + \big|\mathbb{E}\big[e^{iu_2c}\big(e^{iu_1X_n} - e^{iu_1X_\infty}\big)\big]\big| \\ &\leq \big|\mathbb{E}\big[\big|e^{iu_2Y_n} - e^{iu_2c}\big|\big]\big| + \big|\phi_{X_n}(u_1) - \phi_{X_\infty}(u_1)\big|. \end{aligned}$$
Mais $\mathbb{E}\big[\big|e^{iu_2Y_n} - e^{iu_2c}\big|\big] \leq \sqrt{\mathbb{E}\big[\big|e^{iu_2Y_n} - e^{iu_2c}\big|^2\big]} \leq \sqrt{\mathbb{E}\big[2 - e^{iu_2(Y_n-c)} - e^{iu_2(c-Y_n)}\big]} \leq \sqrt{2 - \frac{\phi_{Y_n}(u_2)}{\phi_c(u_2)} - \frac{\phi_{Y_n}(-u_2)}{\phi_c(-u_2)}} \underset{n\to\infty}{\longrightarrow} 0$
car $\phi_{Y_n}(u_2) \underset{n\to\infty}{\longrightarrow} \phi_c(u_2)$ pour tout u_2, puisque $Y_n \xrightarrow[n\to\infty]{\mathcal{P}} c$ donc $Y_n \xrightarrow[n\to\infty]{\mathcal{L}} c$. Par ailleurs, $\big|\phi_{X_n}(u_1) - \phi_{X_\infty}(u_1)\big| \underset{n\to\infty}{\longrightarrow} 0$ pour tout u_1 puisque $X_n \xrightarrow[n\to\infty]{\mathcal{L}} X_\infty$. En conséquence, $\big|\phi_{(X_n,Y_n)}(u) - \phi_{(X_\infty,c)}(u)\big| \underset{n\to\infty}{\longrightarrow} 0$, doù le résultat. □

4.2 Théorèmes limites

Dans la suite, on va s'intéresser à la convergence de la moyenne empirique. On commence par obtenir les propriétés suivantes:

Propriété. *Pour $(X_1, \ldots, X_n)$ une famille de v.a. définies sur $(\Omega, \mathcal{A}, \mathbb{P})$.*

- *Si $\mathbb{E}[|X_k|] < \infty$ pour tout k, alors $\mathbb{E}(\overline{X}_n) = \frac{1}{n}\sum_{k=1}^{n} \mathbb{E}(X_k)$;*
- *Si $\mathbb{E}[X_k^2] < \infty$ pour tout k,*

$$var(\overline{X}_n) = \frac{1}{n^2} \sum_{1 \le i,j \le n} cov(X_i, X_j) = \frac{1}{n^2} \sum_{i=1}^{n} var(X_i) + \frac{2}{n^2} \sum_{1 \le i < j \le n} cov(X_i, X_j).$$

Proof.

- On obtient l'existence de $\mathbb{E}[|\overline{X}_n|]$ par l'inégalité triangulaire et sa valeur par linéarité de l'espérance.
- Comme $\mathrm{var}(\overline{X}_n) = \mathbb{E}[(\overline{X}_n - \mathbb{E}[\overline{X}_n])^2] = \frac{1}{n^2}\sum_{k=1}^{n}(X_k - \mathbb{E}[\overline{X}_n])^2$, on obtient l'existence de $\mathrm{var}(\overline{X}_n)$ par l'inégalité triangulaire pour $\|Z\|_2 = \mathbb{E}[Z^2]^{1/2}$.
  De plus, par bilinéarité de la covariance, on a

$$\mathrm{var}(\overline{X}_n) = \frac{1}{n^2}\mathrm{cov}(\sum_{i=1}^{n} X_i, \sum_{j=1}^{n} X_j) = \frac{1}{n^2}\sum_{1 \le i,j \le n} \mathrm{cov}(X_i, X_j) = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{var}(X_i) + \frac{2}{n^2}\sum_{1 \le i < j \le n} \mathrm{cov}(X_i, X_j).$$

□

Conséquence. Si $(X_1, \ldots, X_n)$ sont des v.a. **indépendantes** et **identiquement distribuées** (notées v.a.i.i.d.) alors

$$\mathbb{E}(\overline{X}_n) = \mathbb{E}[X_1] \quad \text{si } \mathbb{E}[|X_1|] < \infty \quad \text{et} \quad \mathrm{var}(\overline{X}_n) = \frac{1}{n}\mathrm{var}(X_1) \quad \text{si } \mathbb{E}[X_1^2] < \infty.$$

Proof. La preuve est immédiate en utilisant le fait que $\mathbb{E}[X_k] = \mathbb{E}[X_1]$ et $\mathrm{var}(X_k) = \mathrm{var}(X_1)$ pour tout k, et le fait que l'indépendance entre les X_i implique que $\mathrm{cov}(X_i, X_j) = 0$ pour $i \neq j$. □

Propriété (Inégalité de Markov (Tchebychev, 1867, Markov, 1884)). *Soit X une v.a.* **positive** *définie sur $(\Omega, \mathcal{A}, \mathbb{P})$. Alors pour tout $\varepsilon > 0$,*

$$\mathbb{P}(X \ge \varepsilon) \le \frac{\mathbb{E}[X]}{\varepsilon} \quad \textit{(valable si } \mathbb{E}[X| = \infty\textit{)}.$$

Proof. Soit $\varepsilon > 0$. On considère $Y = X\,\mathbb{1}_{X \ge \varepsilon}$. Alors Y est une v.a. positive et $\mathbb{E}[Y] = \mathbb{E}[X\,\mathbb{1}_{X \ge \varepsilon}] \ge \mathbb{E}[\varepsilon\,\mathbb{1}_{X \ge \varepsilon}] = \varepsilon\,\mathbb{P}(X \ge \varepsilon)$. Mais on a également $Y \le X$ par définition, donc $\mathbb{E}[Y] \le \mathbb{E}[X]$. D'où le résultat. □

Voici une conséquence directe de cette inégalité:

Propriété (Inégalité de Bienaymé-Tchebitchev (1853-1867)). *Soit X une v.a. définie sur $(\Omega, \mathcal{A}, \mathbb{P})$ telle que $\mathbb{E}[X^2] < \infty$. Alors $m = \mathbb{E}[X]$ et $\sigma^2 = var(X)$, pour tout $\varepsilon > 0$,*

$$\mathbb{P}\big(|X - \mathbb{E}[X]| \ge \varepsilon\big) \le \frac{var(X)}{\varepsilon^2}.$$

Proof. On reprend l'Inégalité de Markov appliquée à $(X - \mathbb{E}[X])^2$ qui est bien une v.a. positive et à ε^2 plutôt que ε. Le résultat est obtenu puisque $\mathbb{E}[(X - \mathbb{E}[X])^2] = \mathrm{var}(X)$ et en remarquant que $\mathbb{P}\big((X - \mathbb{E}[X])^2 \ge \varepsilon^2\big) = \mathbb{P}\big(|X - \mathbb{E}[X]| \ge \varepsilon\big)$. □

On va maintenant appliquer tout se qui précède à la moyenne empirique d'une famille de variables aléatoires indépendantes ayant toute la même loi:

Théorème (Loi Faible des Grands Nombres avec moment d'ordre 2 (Tchebytchev, 1867)). *Soit* $(X_n)_{n\in\mathbb{N}}$ *une suite de* **v.a.i.i.d.** *définies sur* $(\Omega,\mathcal{A},\mathbb{P})$ *telle que* $\mathbb{E}[X_1^2]<\infty$. *Alors:*

$$\overline{X}_n \xrightarrow[n\to\infty]{\mathcal{P}} \mathbb{E}[X_1].$$

Proof. Application directe de l'inégalité de Bienaymé-Tchebytchev: pour tout $\varepsilon>0$,

$$\mathbb{P}\big(|\overline{X}_n-\mathbb{E}[X_1]|\geq\varepsilon\big)\leq\frac{\mathrm{var}(\overline{X}_n)}{\varepsilon^2}=\frac{\mathrm{var}(X_1)}{n\,\varepsilon^2}\xrightarrow[n\to\infty]{}0,$$

d'où la convergence en probabilité de $\overline{X}_n$. □

Théorème (Loi Faible des Grands Nombres avec moment d'ordre 1 (Khintchine, 1929)). *Soit* $(X_n)_{n\in\mathbb{N}}$ *une suite de* **v.a.i.i.d.** *définies sur* $(\Omega,\mathcal{A},\mathbb{P})$ *telle que* $\mathbb{E}[|X_1|]<\infty$. *Alors:*

$$\overline{X}_n \xrightarrow[n\to\infty]{\mathcal{P}} \mathbb{E}[X_1].$$

Proof. Sans perte de généralité on peut supposer que (X_i) est centré (sinon on considère $X'_k=X_k-\mathbb{E}[X_1]$). Alors pour $u\in\mathbb{R}$, en passant par la fonction caractéristique on:

$$\phi_{\overline{X}_n}(u)=\mathbb{E}\big[e^{i\,\frac{u}{n}\,\sum_{k=1}^n X_k}\big]=\phi^n_{X_1}(u/n),$$

grâce à l'indépendance et le fait d'être identiquement distribué. Mais $|\phi_{X_1}(u/n)|\leq 1$ comme toute fonction caractéristique et on peut effectuer un développement limité de Lagrange d'ordre 1 en 0 de $\phi_{X_1}(u/n)$, sachant que puisque $\mathbb{E}[|X_1|]<\infty$ alors ϕ_{X_1} est de classe $\mathcal{C}^1$ sur $\mathbb{R}$:

$$\phi_{X_1}(u/n)=\phi_{X_1}(0)+\frac{u}{n}\,\phi'\big(\frac{v}{n}\big)=1+\frac{u}{n}\,\phi'\big(\frac{v}{n}\big),$$

avec $|v|\leq|u|$. Mais $\phi'(0)=i\,\mathbb{E}[X_1]=0$, donc par continuité de ϕ' en 0 alors $\phi'\big(\frac{v}{n}\big)\xrightarrow[n\to\infty]{}0$.
On peut alors utiliser le résultat suivant (voir Exercice 9 TD4): si (z_j) et (z'_j) sont deux familles de nombres complexes tels que $|z_j|\leq 1$ et $|z'_j|\leq 1$ pour tout j, alors

$$\Big|\prod_{j=1}^n z_j-\prod_{j=1}^n z'_j\Big|\leq\sum_{j=1}^n|z_j-z'_j|.$$

En prenant ici $z_j=\phi_{X_1}(u/n)$ et $z'_j=1$ pour tout j, on aboutit à:

$$|\phi^n_{X_1}(u/n)-1|\leq\sum_{j=1}^n|1+\frac{u}{n}\,\phi'\big(\frac{v}{n}\big)-1|=u\,\phi'\big(\frac{v}{n}\big)\xrightarrow[n\to\infty]{}0.$$

De ceci on en déduit que pour tout $u\in\mathbb{R}$, $\phi_{\overline{X}_n}(u)\xrightarrow[n\to\infty]{}1$. Or 1 est la fonction caractéristique de la variable aléatoire qui vaut 0. D'où $\overline{X}_n\xrightarrow[n\to\infty]{\mathcal{L}}0$ et comme c'est une convergence vers une constante $\overline{X}_n\xrightarrow[n\to\infty]{\mathcal{P}}0$. □

Théorème (Loi forte des Grands Nombres (Kolmogorov, 1940)). *Soit* $(X_n)_{n\in\mathbb{N}}$ *une suite de* **v.a.i.i.d.** *définies sur* $(\Omega,\mathcal{A},\mathbb{P})$. *Alors*

$$\overline{X}_n \xrightarrow[n\to\infty]{p.s.} m \quad\Longleftrightarrow\quad \mathbb{E}[|X_1|]<\infty \text{ et } m=\mathbb{E}[X_0].$$

Proof. Trop difficile, voir cours de M1 MAEF. □

Il est possible d'être plus précis quant au comportement asymptotique de la moyenne empirique autour de son espérance: c'est ce que précise le théorème suivant, à savoir que ce comportement est gaussien et se resserre à vitesse $1/\sqrt{n}$ autour de l'espérance.

Théorème (Théorème de la limite centrale (Lindeberg et Lévy, $\sim$ 1920)). *Soit* $(X_n)_{n\in\mathbb{N}}$ *une suite de* **v.a.i.i.d.** *définies sur* $(\Omega,\mathcal{A},\mathbb{P})$ *telle que* $\mathbb{E}[X_0^2]<\infty$. *Alors:*

$$\sqrt{n}\Big(\frac{\overline{X}_n-m}{\sigma}\Big)\xrightarrow[n\to\infty]{\mathcal{L}}\mathcal{N}(0,1)\quad \text{où } m=\mathbb{E}[X_0] \text{ et } \sigma^2=var(X_0).$$

Proof. En premier lieu, quitte à considérer $X'_k = (X_k - m)\sigma$, on se place sans perte de généralité dans le cas où les (X_k) sont centrées et de variance 1. On considère alors $Z_n = \sqrt{n}\,\overline{X}_n$. Comme les (X_k) sont des v.a.i.i.d., on a pour tout $u \in \mathbb{R}$ (voir preuve précédente),

$$\phi_{Z_n}(u) = \mathbb{E}\big[e^{\,i\,\frac{u}{\sqrt{n}}\,\sum_{k=1}^n X_k}\big] = \phi^n_{X_1}(u/\sqrt{n}).$$

Comme $\mathbb{E}[X_1^2] < \infty$, alors ϕ_{X_1} est de classe $\mathcal{C}^2$ sur $\mathbb{R}$, et un développement limité d'ordre 2 en 0 de $\phi_{X_1}(u/\sqrt{n})$ donne:

$$\phi_{X_1}(u/\sqrt{n}) = \phi_{X_1}(0) + \phi'_{X_1}(0)\,\frac{u}{\sqrt{n}} + \phi''_{X_1}(0)\,\frac{u^2}{2\,n}\big(1+\varepsilon_n\big) = 1 - \frac{u^2}{2\,n}\big(1+\varepsilon_n\big),$$

car $\phi'_{X_1}(0) = i\,\mathbb{E}[X_1] = 0$ et $\phi''_{X_1}(0) = -\mathbb{E}[X_1^2] = -1$. Comme dans la preuve précédente on utilise le fait $\Big|\prod_{j=1}^n z_j - \prod_{j=1}^n z'_j\Big| \leq \sum_{j=1}^n |z_j - z'_j|$ avec cette fois-ci $z_j = \phi_{X_1}(u/\sqrt{n})$ et $z'_j = 1 - \frac{u^2}{2\,n}$, et on obtient que:

$$|\phi^n_{X_1}(u/\sqrt{n}) - \big(1 - \frac{u^2}{2\,n}\big)^n| \leq \sum_{j=1}^n \frac{u^2}{2\,n}\,|\varepsilon_n| \leq \frac{u^2}{2}\,|\varepsilon_n| \underset{n\to\infty}{\longrightarrow} 0.$$

Or $\big(1 - \frac{u^2}{2\,n}\big)^n = e^{\,n\,\ln\left((1-\frac{u^2}{2\,n}\right)} \underset{n\to\infty}{\longrightarrow} e^{-u^2/2}$. On en déduit donc que $\phi_{Z_n}(u) \underset{n\to\infty}{\longrightarrow} e^{-u^2/2}$ pour tout $u \in \mathbb{R}$, et $e^{-u^2/2}$ est la fonction caractéristique de la loi $\mathcal{N}(0,1)$. □

Théorème (Second théorème de la limite centrale). *Soit $(X_n)_{n\in\mathbb{N}}$ une suite de* **v.a.i.i.d.** *définies sur $(\Omega, \mathcal{A}, \mathbb{P})$ telle que $\mathbb{E}[X_0^2] < \infty$. Alors :*

$$\sqrt{n}\Big(\frac{\overline{X}_n - m}{\widehat{\sigma}_n}\Big) \underset{n\to\infty}{\overset{\mathcal{L}}{\longrightarrow}} \mathcal{N}(0,1) \quad \textit{avec } m = \mathbb{E}[X_0] \textit{ et } \widehat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)^2.$$

Proof. On montre d'abord que $\widehat{\sigma}_n^2 \underset{n\to\infty}{\overset{\mathcal{P}}{\longrightarrow}} \sigma^2$ (on utilise la LGN pour $\frac{1}{n}\sum_{k=1}^n X_k^2$, et comme $\overline{X}_n \underset{n\to\infty}{\overset{\mathcal{P}}{\longrightarrow}} m$, donc $\overline{X}_n^2 \underset{n\to\infty}{\overset{\mathcal{P}}{\longrightarrow}} m^2$ on utilise le Lemme de Slutsky pour montrer que $\widehat{\sigma}_n^2 = \frac{1}{n}\sum_{k=1}^n X_k^2 - \overline{X}_n^2 \underset{n\to\infty}{\overset{\mathcal{P}}{\longrightarrow}} \sigma^2$). Donc $\frac{\sigma}{\widehat{\sigma}_n} \underset{n\to\infty}{\overset{\mathcal{P}}{\longrightarrow}} 1$ (fonction continue). Or

$$\sqrt{n}\Big(\frac{\overline{X}_n - m}{\widehat{\sigma}_n}\Big) = \frac{\sigma}{\widehat{\sigma}_n}\,\sqrt{n}\Big(\frac{\overline{X}_n - m}{\sigma}\Big).$$

En utilisant le Lemme de Slutsky, comme $\frac{\sigma}{\widehat{\sigma}_n} \underset{n\to\infty}{\overset{\mathcal{P}}{\longrightarrow}} 1$ et $\sqrt{n}\Big(\frac{\overline{X}_n - m}{\sigma}\Big) \underset{n\to\infty}{\overset{\mathcal{L}}{\longrightarrow}} \mathcal{N}(0,1)$, on obtient le résultat demandé. □

Exemple: Application de ce TLC pour les v.a. de Bernoulli de paramètre p.

Remarque: Dans le cas où les v.a. sont gaussiennes, alors on a pour tout $n \geq 2$,

$$\sqrt{n}\Big(\frac{\overline{X}_n - m}{\widehat{\sigma}_n}\Big) \overset{\mathcal{L}}{\sim} t(n-1).$$

On en déduit donc également que $t(n-1) \underset{n\to\infty}{\overset{\mathcal{L}}{\longrightarrow}} \mathcal{N}(0,1)$.

Un dernier théorème limite, souvent utile en statistique, peut être énoncé. Il s'apparente à une formule de Taylor:

Théorème (Delta-méthode (Kelley, 1928). *Soit $(Z_n)_{n\in\mathbb{N}}$ une suite de v.a. définies sur $(\Omega, \mathcal{A}, \mathbb{P})$ et telle que $a_n\,(Z_n - m) \underset{n\to\infty}{\overset{\mathcal{L}}{\longrightarrow}} \mathcal{N}(0,1)$, avec $a_n \underset{n\to\infty}{\longrightarrow} \infty$ et $m \in \mathbb{R}$. Alors, pour toute fonction g de classe $\mathcal{C}^1$ dans un voisinage de m telle que $g'(m) \neq 0$,*

$$a_n\,\big(g(Z_n) - g(m)\big) \underset{n\to\infty}{\overset{\mathcal{L}}{\longrightarrow}} \mathcal{N}\big(0\,,\,(g'(m))^2\big).$$

Proof. D'après le développement de Taylor-Lagrange, il existe une variable aléatoire λ à valeurs dans $[0,1]$ telle que

$$a_n\,\big(g(Z_n) - g(m)\big) = g'\big(\lambda\,Z_n + (1-\lambda)\,m\big)\,a_n\,(Z_n - m).$$

Le fait que l'on ait $a_n\,(Z_n - m) \underset{n\to\infty}{\overset{\mathcal{L}}{\longrightarrow}} \mathcal{N}(0,1)$ implique que $Z_n \underset{n\to\infty}{\overset{\mathcal{P}}{\longrightarrow}} m$ et donc $\lambda\,Z_n + (1-\lambda)\,m \underset{n\to\infty}{\overset{\mathcal{P}}{\longrightarrow}} \lambda\,m + (1-\lambda)\,m = m$. Comme g' est supposée être une fonction continue implique que $g'\big(\lambda\,Z_n + (1-\lambda)\,m\big) \underset{n\to\infty}{\overset{\mathcal{P}}{\longrightarrow}} g'(m)$. Enfin, grâce au Lemme de Slutsky, $a_n\,\big(g(Z_n) - g(m)\big) \underset{n\to\infty}{\overset{\mathcal{L}}{\longrightarrow}} g'(m)\,\mathcal{N}(0\,,\,1)$. □

Exemple d'application: Si on a le TLC pour $\overline{X}_n$ et si g de classe $\mathcal{C}^1$ dans un voisinage $\mathbb{E}[X_0]$, alors:

$$\sqrt{n}\left(g\big(\overline{X}_n\big) - g\big(\mathbb{E}[X_0]\big)\right) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}\big(0\,,\, \big(g'\big(\mathbb{E}[X_0]\big)\big)^2 \operatorname{var}(X_0)\big).$$

A utiliser notamment pour obtenir un TLC pour $1/\overline{X}_n$ dans le cas de v.a.i.i.d. de loi exponentielle de paramètre λ, ce qui fournit un TLC pour l'estimateur par maximum de vraisemblance de λ.

5 Estimation paramétrique

5.1 Statistique, modèle statistique, définitions et premières propriétés

Définition. Si on dispose de $(X_1, \ldots, X_n)$ une famille de n vecteurs aléatoires définies sur le même espace de probabilités $(\Omega, \mathcal{A}, \mathbb{P})$, une **statistique** $\widehat{T}$ de $(X_1, \ldots, X_n)$ est une **fonction mesurable** h de $(X_1, \ldots, X_n)$, soit $\widehat{T} = h(X_1, \ldots, X_n)$.

Attention! Dans la suite, nous allons parler d'un type de statistique, les estimateurs. Mais une statistique peut aussi être une statistique de test, une statistique de classification, une statistique de prédiction,...

Définition. Si on dispose de $(X_1, \ldots, X_n)$ une famille de n variables aléatoires définies sur le même espace de probabilités $(\Omega, \mathcal{A}, \mathbb{P})$, un **estimateur** d'un vecteur $\theta \in \mathbb{R}^d$ est une **fonction mesurable** de $(X_1, \ldots, X_n)$ **ne dépendant pas de** θ.

Cette définition est très générale. Un exemple paradigmatique va nous aider à la comprendre:

Exemple de l'estimation d'une moyenne (espérance) de v.a.i.i.d.:
Considérons un échantillon observé $(X_1, \ldots, X_n)$ issu d'une suite (X_k) de v.a.i.i.d. Nous supposerons également que $\mathbb{E}[X_0^2] < \infty$ et que l'on désire, à partir de $(X_1, \ldots, X_n)$ estimer $m = \mathbb{E}[X_0]$. Plus concrétement, à partir de l'observation de la température de 150 mois de janvier à Paris, on pourrait se demander qu'elle est la température moyenne en janvier à Paris. Dans un tel cadre, on pourrait dénommer X_i.
Soit $(X_n)_{n\in\mathbb{N}}$ une suite de variables aléatoires **indépendantes** définies sur le même espace de probabilités $(\Omega, \mathcal{A}, \mathbb{P})$ et **identiquement distribuées** par rapport à une loi dont l'espérance $m \in \mathbb{R}$ existe mais est **inconnue** (on suppose également que $\sigma^2 < \infty$). On suppose que l'on a **observé** $(X_1, \ldots, X_n)$, c'est à dire qu'il existe $\omega \in \Omega$ tel que l'on connaisse le vecteur $(X_1(\omega), \ldots, X_n(\omega))$.
On a vu précédemment que $\overline{X}_n \xrightarrow[n\to\infty]{p.s.} m$, ce qui signifie que pour n "grand" et pour presque tout $\omega \in \Omega$, $X_n(\omega)$ sera "proche" de m. Ainsi $\overline{X}_n$ qui est bien une fonction de $(X_1, \ldots, X_n)$ a légitimité à estimer m.
Est-ce le seul estimateur possible? Non. On aurait pu choisir aussi $\overline{X}_{n-1}$, ou plus généralement $\sum_{k=1}^n p_k X_k$ en ayant choisi des poids p_k tels que $\sum_{k=1}^n p_k = 1$. Pour certaines lois, il pourra aussi être intéressant d'estimer m par la médiane empirique de $(X_1, \ldots, X_n)$ ou bien par $\frac{1}{2}\big(\max_{1\le k\le n}(X_k) - \min_{1\le k\le n}(X_k)\big)$, qui sont également des fonctions mesurables de $(X_1, \ldots, X_n)$. Mais en revanche, on ne pouvait considérer comme fonction $f(X_1, \ldots, X_n) = m$, qui est bien mesurable mais qui dépend de m (dommage, on ne pouvait faire mieux pour estimer m, mais on conçoit bien que ce n'est aucunement réaliste...).
On va maintenant essayer de mettre un peu "d'ordre" dans ce choix d'estimateur, et pour cela nous allons d'abord nous placer dans un cadre bien délimité, celui des modèles statistiques paramétriques.

Dans toute la suite, on se place sur $(\Omega, \mathcal{A}, \mathbb{P})$ un espace de probabilité. On considère $(X_n)_{n\in\mathbb{N}}$ une suite de variable aléatoire, où chaque X_i est définie sur $(\Omega, \mathcal{A}, \mathbb{P})$ et est à valeur dans $\Omega' \subset \mathbb{R}$.

Définition. On appelle **modèle statistique** de dimension n un espace $((\Omega')^n, \mathcal{A}'_n, \mu_n)$, où $\mathcal{A}'_n$ est une tribu sur $(\Omega')^n$ et μ_n une mesure de probabilité sur $((\Omega')^n, \mathcal{A}'_n)$. Un **échantillon** de taille n du modèle statistique $((\Omega')^n, \mathcal{A}'_n, \mu)$ est un vecteur aléatoire $(X_1, \ldots, X_n)$ distribuée selon la loi μ_n. Pour $\omega \in \Omega$, $(X_1(\omega), \ldots, X_n(\omega))$ vecteur de $\mathbb{R}^n$ est appelé **échantillon observé**.

Définition. On appelle:

- **Modèle statistique paramétrique**, une famille de modèle de la forme: $((\Omega')^n, \mathcal{A}'_n, \mathbb{P}_\theta^{(n)}, \theta \in \Theta)$, où $\Theta \subset \mathbb{R}^p$.
- **Modèle statistique semi-paramétrique**, une famille de modèle de la forme: $((\Omega')^n, \mathcal{A}'_n, \mathbb{P}_{(\theta,f)}^{(n)}, \theta \in \Theta, f \in \mathcal{F})$, où $\Theta \subset \mathbb{R}^p$ et $\mathcal{F}$ n'est pas de dimension finie.
- **Modèle statistique non-paramétrique**, une famille de modèle de la forme: $((\Omega')^n, \mathcal{A}'_n, \mathbb{P}_f^{(n)}, f \in \mathcal{F})$, où $\mathcal{F}$ n'est pas de dimension finie.

Définition. Soit le modèle paramétrique $((\Omega')^n, \mathcal{A}'_n, \mathbb{P}_\theta^{(n)}, \theta \in \Theta)$ où $\Theta \subset \mathbb{R}^p$.

- On dit que ce modèle est **dominé** par une mesure μ_n lorsque $\mathbb{P}_\theta^{(n)}$ est absolument continue par rapport à μ_n pour tout $\theta \in \Theta$.
- Si ce modèle est dominée par μ_n, on appelle **vraisemblance du modèle** la fonction

$$\theta \in \Theta \mapsto L_\theta(x_1, \cdots, x_n) = \frac{\partial \mathbb{P}_\theta^{(n)}}{\partial \mu_n}(x_1, \cdots, x_n) \quad \text{pour } (x_1, \cdots, x_n) \in (\Omega')^n.$$

Exemples:

- Dans le cas où μ_n est la mesure de Lebesgue sur $\mathbb{R}^n$, la vraisemblance sera la densité (classique) en $(x_1, \cdots, x_n)$.
- Dans le cas où μ_n est la mesure de comptage sur $\mathbb{N}^n$, la vraisemblance sera la probabilité en $(x_1, \cdots, x_n)$.
- Attention! si le support de $\mathbb{P}_\theta^{(n)}$ dépend de θ, la mesure qui domine (ainsi que Ω' et $\mathcal{A}'_n$) ne peut dépendre de θ: il ne faut pas oublier de le préciser dans l'expression de la vraisemblance.

Pour mesurer l'information fournie par un modèle paramétrique dominé (ou une statistique sur ce modèle) au sujet d'un paramètre, une idée naturelle serait de mesurer comment varie localement la mesure de probabilité, ou encore sa vraisemblance. Les fluctuations moyennes de cette vraisemblance seront donc un bon indicateur: pour ce faire on considérera, lorsqu'il existe $\mathrm{grad}_\theta(L_\theta(X_1, \cdots, X_n))$, et on s'intéressera à la matrice de covariance de $\mathrm{grad}_\theta(L_\theta(X_1, \cdots, X_n))$, dont on peut montrer qu'elle ne dépend pas du choix de la mesure dominante choisie. Précisons d'abord la notion de modèle régulier qui nous permettra de définir cette quantité d'information.

Définition. Dans le cadre d'un modèle statistique paramétrique $((\Omega')^n, \mathcal{A}'_n, \mathbb{P}_\theta, \theta \in \Theta)$, où $\Theta \subset \mathbb{R}^p$, dominé par une mesure μ, on dira que ce modèle est **régulier** lorsque:

1. Θ est un ouvert de $\mathbb{R}^d$;
2. la vraisemblance $L_\theta(.)$ vérifie $\forall (x_1, \ldots, x_n) \in (\Omega')^n$, $\forall \theta \in \Theta$, $L_\theta(x_1, \ldots, x_n) > 0$;
3. $\forall (x_1, \ldots, x_n) \in (\Omega')^n$, la fonction $\theta \in \Theta \mapsto \log(L_\theta(.))$ est différentiable sur Θ par rapport à θ, et son gradient appartient à $\mathbb{L}^2((\Omega')^n, \mathcal{A}'_n, \mathbb{P}_\theta)$ $\forall \theta \in \Theta$;

4. $\forall \theta \in \Theta$, pour toute fonction $h : \mathbb{R}^n \to \mathbb{R}$ appartenant à $\mathbb{L}^1((\Omega')^n, \mathcal{A}'_n, \mathbb{P}_\theta)$, alors :

$$\frac{\partial}{\partial\theta}\int_{(\Omega')^n} h(x)\cdot L_\theta(x)\,d\mu(x) = \int_{(\Omega')^n} h(x)\cdot \frac{\partial}{\partial\theta}L_\theta(x)\,d\mu(x). \tag{1}$$

Exemples: Lorsque le modèle statistique est issu d'une famille de v.a.i.i.d., alors les lois de type Bernoulli, binomiale, géométrique, Poisson, exponentielle, gamma ou gaussienne induisent des modèles réguliers lorsque ce sont leurs paramètres que l'on considère. En revanche, une loi comme la loi uniforme pour laquelle on veut estimer les paramètres ne définit pas un modèle régulier.

Propriété. *Pour un modèle régulier,* $\mathbb{E}_\theta\big[grad_\theta(\log L_\theta(.))\big] = 0$.

Proof. On a $\mathbb{E}_\mu[L_\theta(.)] = 1$ donc $\mathbb{E}_\mu[\mathrm{grad}_\theta L_\theta(.)] = 0$. Ainsi, $\mathbb{E}_\theta\big[\frac{\mathrm{grad}_\theta(L_\theta(.))}{L_\theta(.)}\big] = 0$, soit $\mathbb{E}_\theta\big[\mathrm{grad}_\theta(\log L_\theta(.))\big] = 0$. □

Définition. Pour un modèle statistique paramétrique dominé régulier, on appelle **information de Fisher**, la matrice:

$$I_n(\theta) = \Big(\mathbb{E}_\theta\Big[\frac{\partial(\log L_\theta(X_1,\cdots,X_N))}{\partial\theta_i}\times\frac{\partial(\log L_\theta(X_1,\cdots,X_N))}{\partial\theta_j}\Big]\Big)_{1\le i,j\le p}.$$

Propriété. *Pour un modèle statistique paramétrique dominé régulier, et si* $\forall (x_1,\ldots,x_n)\in(\Omega')^n$, *la fonction* $\theta\in\Theta\mapsto\log(L_\theta(.))$ *est* $\mathcal{C}^2(\Theta)$, *alors:*

$$I_n(\theta) = -\Big(\mathbb{E}_\theta\Big[\frac{\partial^2(\log L_\theta(X_1,\cdots,X_N))}{\partial\theta_i\,\partial\theta_j}\Big]\Big)_{1\le i,j\le p}.$$

Proof. En reprenant la preuve précédente, $\mathbb{E}_\mu[L_\theta(.)] = 1$ donc $\mathbb{E}_\mu\big[\frac{\partial^2}{\partial\theta_i\partial\theta_j}L_\theta(.)\big] = 0$, d'où $\mathbb{E}_\theta\big[\frac{1}{L_\theta(.)}\frac{\partial^2}{\partial\theta_i\partial\theta_j}L_\theta(.)\big] = 0$. Mais

$$\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log\left(L_\theta(.)\right) = \frac{1}{L_\theta(.)}\frac{\partial^2}{\partial\theta_i\partial\theta_j}L_\theta(.) - \frac{1}{L_\theta^2(.)}\frac{\partial}{\partial\theta_i}L_\theta(.)\frac{\partial}{\partial\theta_j}L_\theta(.)\Big) = \frac{1}{L_\theta(.)}\frac{\partial^2}{\partial\theta_i\partial\theta_j}L_\theta(.) - \frac{\partial}{\partial\theta_i}\log\left(L_\theta(.)\right)\frac{\partial}{\partial\theta_j}\log\left(L_\theta(.)\right).$$

En considérant $\mathbb{E}_\theta$ de cette expression, on obtient bien le résultat demandé. □

Conséquence: Sous les mémes hypothèses, et si le modèle est constituée de v.a.i.i.d., alors:

$$I_n(\theta) = -n\,\Big(\mathbb{E}_\theta\Big[\frac{\partial^2(\log L_\theta(X_1))}{\partial\theta_i\,\partial\theta_j}\Big]\Big)_{1\le i,j\le p}.$$

5.2 Estimation paramétrique: cadre général

On se place dans le cadre d'un modèle statistique paramétrique $((\Omega')^n, \mathcal{A}'_n, \mathbb{P}_\theta^{(n)}, \theta\in\Theta)$, où $\Theta\subset\mathbb{R}^p$, dominé par une mesure μ. Par ailleurs, on suppose que Θ est un ouvert.

Définition.

- S'il existe, on appelle **biais** d'un estimateur $\widehat{T}$ de θ vecteur de $\mathbb{R}^p$, $B(\theta) = \mathbb{E}_\theta[\widehat{T}] - \theta$. On dira que l'estimateur est sans biais si $B(\theta) = 0$ pour tout $\theta\in\Theta$.
- S'il existe, on appelle **risque quadratique** de l'estimateur $\widehat{T}$ de θ le réel positif $R(\theta) = \mathbb{E}_\theta\big[\|\widehat{T}-\theta\|^2\big]$, où $\|.\|$ désigne usuellement la norme euclidienne (mais peut être une autre fonctionnelle positive et convexe). Si l'estimateur est sans biais alors, $R(\theta) = \mathrm{Trace}(\mathrm{cov}(\widehat{T}))$.

Propriété. *Sous les hypothèses précédentes, on a* $R(\theta) = \big\|B(\theta)\big\|^2 + Trace(cov(\widehat{T}))$ *pour tout* $\theta\in\Theta$.

Pour pouvoir parler du comportement asymptotique d'une statistique, on va devoir se placer dans un "gros" modèle, dans lequel un échantillon est une suite de v.a. En quelque sorte, ce gros modèle pourra s'écrire $((\Omega')^{\mathbb{N}}, \mathcal{A}'^{\mathbb{N}}_{\mathbb{N}}, \mathbb{P}_\theta^{\mathbb{N}}, \theta\in\Theta)$, où $\Theta\subset\mathbb{R}^p$ (la dimension du paramètre reste constante). Pour un n fixé, une statistique $\widehat{T}_n$ sera d'abord une projection du "gros" modèle sur le modèle de taille n, puis une statistique "normale". On devra donc parler d'une suite d'estimateurs $(\widehat{T}_n)_n$

Définition. Pour un modèle statistique paramétrique $((\Omega')^{\mathbb{N}}, \mathcal{A}'_{\mathbb{N}}, \mathbb{P}_\theta^{\mathbb{N}}, \theta \in \Theta)$, où $\Theta \subset \mathbb{R}^p$, et pour $(\widehat{T}_n)_n$ une suite d'estimateurs de θ:

- Avec B_n le biais de $\widehat{T}_n$, si $\lim_{n\to\infty} B_n(\theta) = 0$, on dit que la suite d'estimateur $(\widehat{T}_n)$ est **asymptotiquement sans biais**.
- On dit que $(\widehat{T}_n)_n$ est **convergent** lorsque $\widehat{T}_n \xrightarrow[n\to\infty]{\mathcal{P}} \theta$.
- S'il existe (a_n) une suite de réels positifs $\to +\infty$ telle que $a_n(\widehat{T}_n - \theta) \xrightarrow[n\to\infty]{\mathcal{L}} Z_\theta$, où $Z_\theta \neq 0$ (ne dépendant pas de n!), on dit $(\widehat{T}_n)_n$ converge vers θ à la **vitesse** a_n.

A priori, être sans biais n'est pas un bon critère pour garantir une certaine optimalité de la convergence d'un estimateur. D'ailleurs, l'estimateur peut être non biaisé mais ne pas être convergent (par exemple, si $(X_1, \ldots, X_n)$ sont des v.a.i.i.d. de loi $\mathcal{B}(p)$ avec $0 < p < 1$, alors $\widehat{T}_n = X_1$ est un estimateur non biaisé mais non convergent de p). On préférera plutôt discriminer entre de potentiels estimateurs à l'aide d'un critère portant sur le risque quadratique ou sur la matrice de variance-covariance. Cependant, il n'existe pas de résultats généraux pour trouver un "meilleur" estimateur en ce sens. Pour en obtenir, on devra se limiter à une certaine classe d'estimateurs, celle des estimateurs sans biais.

Définition. Soit un modèle statistique paramétrique $((\Omega')^n, \mathcal{A}'_n, \mathbb{P}_\theta^{(n)}, \theta \in \Theta)$, et soit $\widehat{T}$ un estimateur sans biais de θ. On dit que $\widehat{T}$ est de **variance uniformément minimum parmi les estimateurs sans biais de** θ lorsque pour tout estimateur $\widehat{S}$ sans biais de θ, on a $\forall \theta \in \theta$, $\mathrm{cov}(\widehat{T}) \leq \mathrm{cov}(\widehat{S})$ (au sens où $\mathrm{cov}(\widehat{S}) - \mathrm{cov}(\widehat{T})$ est une matrice positive).

Propriété. *Si $\widehat{T}$ est un estimateur de variance uniformément minimum parmi les estimateurs sans biais, alors il est unique $\mathbb{P}_\theta^{(n)}$-p.s.*

Proof. Soit $\widehat{S}$ un autre estimateur que l'on suppose également de variance uniformément minimum parmi les estimateurs sans biais. Montrons d'abord que $E_\theta\big[(\widehat{T} - \widehat{S}) \cdot {}^t\widehat{T}\big] = 0$. En effet, si $\alpha \in \mathbb{R}$, comme $\widehat{T}$ est de variance minimum, en utilisant des inégalités sur les matrices symétriques:

$$\begin{array}{rcl} \mathrm{cov}(\widehat{T}) & \leq & \mathrm{cov}(\widehat{T} + \alpha(\widehat{T} - \widehat{S})) \\ & \leq & \mathrm{cov}(\widehat{T}) + \alpha^2 \,\mathrm{cov}(\widehat{T} - \widehat{S}) + 2\alpha\, \mathbb{E}_\theta[\widehat{T}\,{}^t(\widehat{T} - \widehat{S})] \\ \Longrightarrow \quad 0 & \leq & \alpha\left(\alpha\,\mathrm{cov}(\widehat{T} - \widehat{S}) + 2\,\mathbb{E}_\theta[\widehat{T} \cdot {}^t(\widehat{T} - \widehat{S})]\right) \quad \text{pour tout } \alpha \in \mathbb{R}. \end{array}$$

Comme $\mathrm{cov}(\widehat{T} - \widehat{S})$ est une matrice positive, la seule possibilité pour avoir la dernière inégalité est que: $\mathbb{E}_\theta\big[\widehat{T} \cdot {}^t(\widehat{T} - \widehat{S})\big] = 0$. Par suite, comme

$$\mathrm{cov}(\widehat{T} - \widehat{S}) = \mathbb{E}_\theta\big[(\widehat{T} - \widehat{S}) \cdot {}^t(\widehat{T} - \widehat{S})\big] = \mathbb{E}_\theta\big[\widehat{T} \cdot {}^t(\widehat{T} - \widehat{S})\big] - \mathbb{E}_\theta\big[\widehat{S} \cdot {}^t(\widehat{T} - \widehat{S})\big],$$

et que l'on a supposé $\widehat{T}$ et $\widehat{S}$ de variance minimum, $\mathrm{cov}(\widehat{T} - \widehat{S}) = 0$. Donc $\widehat{T} = \widehat{S}$ sur un ensemble de $\mathbb{P}_\theta$-mesure égale à 1. □

On aimerait maintenant connaître un peu mieux la covariance d'un tel estimateur. Il est possible d'avoir un résultat très précis lorsque le modèle est régulier:

Théorème (Inégalité de Cramer-Rao (Aitken et Silverstone, 1942)). *Soit un modèle statistique paramétrique $((\Omega')^n, \mathcal{A}'_n, \mathbb{P}_\theta^{(n)}, \theta \in \Theta)$ dominé et* **régulier**, *et soit $\widehat{T}$ un estimateur sans biais de θ, tel que $\mathbb{E}_\theta\big[\|\widehat{T}\|^2\big] < \infty$. Si on suppose que pour tout $\theta \in \Theta$ l'information de Fisher $I_n(\theta)$ est une matrice définie positive, alors, pour tout $\theta \in \Theta$:*

$$cov(\widehat{T}) \geq (I_n(\theta))^{-1} \quad (\textit{au sens des matrices symétriques}).$$

Proof. Soit $Z_\theta(x) = \mathrm{grad}\big(\log(L_\theta(x))\big)$ où $x \in (\Omega')^n$ suit $\mathbb{P}_\theta^{(n)}$. On sait que comme le modèle est régulier, $\mathbb{E}_\theta[Z_\theta] = 0$ pour tout $\theta \in \Theta$ et donc:

$$\mathrm{cov}(Z_\theta) = I_n(\theta) \quad \text{pour tout } \theta \in \Theta.$$

De plus, $\widehat{T}$ est un estimateur sans biais de θ donc pour tout $\theta \in \Theta$:

$$\begin{aligned}
\mathbb{E}_\theta[\widehat{T}] = \theta \quad &\Longrightarrow \quad \int_{(\Omega')^n} \widehat{T}(x)\,{}^t\big(\frac{\partial L_\theta}{\partial \theta}(x)\big)\, d\mu_n(x) = I_p \ \text{ (en dérivant et avec } I_p \text{ la matrice identité de taille } p) \\
&\Longrightarrow \quad \int_{(\Omega')^n} \widehat{T}(x)\,{}^t\big(\frac{\partial L_\theta}{\partial \theta}(x)\big)\, (L_\theta(x))^{-1}\, d\mathbb{P}_\theta^{(n)}(x) = I_p \\
&\Longrightarrow \quad \mathbb{E}_\theta\big[\widehat{T}\,{}^t Z_\theta\big] = \mathrm{cov}_\theta\big(\widehat{T}, Z_\theta\big) = I_p.
\end{aligned}$$

Ainsi, d'après ce qui précède,

$$\begin{aligned}
\mathrm{cov}_\theta(\widehat{T} - I_n^{-1}(\theta)\cdot Z_\theta) &= \mathrm{cov}_\theta(\widehat{T}) - 2\, I_n^{-1}(\theta) + I_n^{-1}(\theta) \\
&= \mathrm{cov}_\theta(\widehat{T}) - I_n^{-1}(\theta).
\end{aligned}$$

En conséquence, comme $\mathrm{cov}_\theta(\widehat{T} - I_n^{-1}(\theta)\, Z_\theta)$ est une matrice positive, l'inégalité de Cramer-Rao est prouvée. □

Corollaire. *Deux cas particuliers méritent attention :*

- *Si le modèle est de la forme $((\Omega')^n, \mathcal{A}'_n, (f_\theta \cdot d\mu)^{\otimes n}, \theta \in \Theta)$, alors $I_n(\theta) = n \cdot I_1(\theta)$, où $I_1(\theta)$ est la matrice d'information de Fisher d'une seule variable aléatoire X distribuée suivant $f_\theta \cdot d\mu$ et l'Inégalité de Cramer-Rao devient donc :*

  $$cov(\widehat{T}) \geq \frac{1}{n}\,(I_1(\theta))^{-1} \quad (\textit{au sens des matrices symétriques}).$$

  On voit donc que pour un échantillon de variables indépendantes et identiquement distribuées, si la vraisemblance est régulière, alors la vitesse de convergence de tout estimateur sans biais est au mieux en $\sqrt{n}$.

- *Si le modèle n'est pas régulier, mais que sous la probabilité $\mathbb{P}_\theta^{(n)}$, la matrice d'information de Fisher existe et est inversible, et surtout si la propriété (1) est vérifiée, alors l'Inégalité de Cramer-Rao est vérifiée.* **Cela exclut cependant les modèles dont le support de $\mathbb{P}_\theta^{(n)}$ dépend de θ,** *comme par exemple le simple modèle de v.a.i.i.d. de loi $\mathcal{U}(]0,\theta[)$, avec $\theta > 0$.*

Définition. Si un estimateur sans biais atteint (respectivement asymptotiquement) la borne de Cramer-Rao (qui ne dépend pas de l'estimateur), on dit qu'il est (resp. asymptotiquement) **efficace**.

Remarque: Un estimateur peut être sans biais, de variance minimale, mais ne pas atteindre la borne de Cramer-Rao, donc ne pas être efficace. De la même manière, il peut exister des estimateurs biaisés atteignant la borne de Cramer-Rao.

5.3 Modèle statistique exponentiel (hors programme)

Définition. On suppose un modèle paramétrique $((\Omega')^n, \mathcal{A}'_n, \mathbb{P}_\theta^{(n)}, \theta \in \Theta \subset \mathbb{R}^p)$ dominé par une mesure μ_n. Si, pour tout $(x_1, \cdots, x_n) \in (\Omega')^n$ et $\theta \in \Theta$, la vraisemblance de ce modèle par rapport à μ_n peut s'écrire sous la forme:

$$L_\theta(x_1, \cdots, x_n) = \exp\Big(\beta(\theta) + b(x_1, \cdots, x_n) + \sum_{j=1}^{p} a_j(x_1, \cdots, x_n)\cdot \alpha_j(\theta)\Big), \tag{2}$$

avec les fonctions $a_j : (\Omega')^n \to \mathbb{R}$, $b : (\Omega')^n \to \mathbb{R}$, $\alpha_j : \Theta \subset \mathbb{R}^p \to \mathbb{R}$, et $\beta : \Theta \to \mathbb{R}$, alors on dit que le modèle est **exponentiel** (ou qu'il appartient à la **famille exponentielle**).

Exemples: Appartiennent à la famille exponentielle les lois:

- Loi discrètes: Lois de Bernoulli, binomiales, de Poisson,...
- Loi "continues": Lois normales, exponentielles, gamma, du chi-deux,...

Remarque: Si $(X_1, \cdots, X_n)$ est un n-échantillon d'un modèle exponentiel (avec θ fixé) alors l'ensemble des valeurs prises par $(X_1, \cdots, X_n)$ ne dépend pas du paramètre θ. Donc, par exemple, le modèle $\left([0,\infty[^n, \mathcal{B}([0,\infty[^n), \left(\mathcal{U}([0,\theta]^n\right)_{\theta>0}\right)$ n'est pas un modèle exponentiel.

Nous allons voir que 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.

Théorème. *Soit un modèle statistique paramétrique $((\Omega')^n, \mathcal{A}'_n, \mathbb{P}_\theta^{(n)}, \theta \in \Theta)$, avec $\Theta \subset \mathbb{R}^p$, dominé et régulier. Alors $\widehat{T} = {}^t(\widehat{T}_1, \ldots, \widehat{T}_p)$ est un estimateur sans biais de θ atteignant la borne de Cramer-Rao si et seulement si le modèle est exponentiel et plus précisément s'il existe des fonctions $a : (\Omega')^n \to \mathbb{R}$, $\beta : \Theta \to \mathbb{R}$ et $\alpha_j : \Theta \to \mathbb{R}$ $(1 \leq j \leq p)$, telles que pour tout $\theta \in \Theta$, $\theta = -\left(\frac{\partial \alpha_j}{\partial \theta_i}(\theta)\right)^{-1}_{1\leq i,j\leq p} \cdot \frac{\partial \beta}{\partial \theta}(\theta)$ et*

$$L_\theta(x_1, \cdots, x_n) = \exp\Big(\beta(\theta) + b(x_1, \cdots, x_n) + \sum_{j=1}^{p} T_j(x_1, \cdots, x_n) \cdot \alpha_j(\theta)\Big).$$

Proof. $\Longleftarrow$ On suppose donc le modèle exponentiel décrit dans le théorème. Si on dérive par rapport à θ un tel modèle, on obtient que pour μ-presque tout $x \in (\Omega')^n$:

$$\frac{\partial}{\partial \theta}(\log L_\theta(x)) = \left(\frac{\partial \alpha_j}{\partial \theta_i}(\theta)\right)_{1\leq i,j\leq p} \cdot \widehat{T} + \frac{\partial \beta}{\partial \theta}(\theta), \quad \text{pour tout } \theta \in \Theta. \tag{3}$$

En conséquence, comme $I(\theta) = \mathbb{E}_\theta\left(\left(\frac{\partial}{\partial \theta}(\log L_\theta(.))\right) \cdot {}^t\left(\frac{\partial}{\partial \theta}(\log L_\theta(.))\right)\right)$, on en déduit que :

$$I(\theta) = \left(\frac{\partial \alpha_j}{\partial \theta_i}(\theta)\right) \cdot \operatorname{cov}_\theta(\widehat{T}) \cdot {}^t\left(\frac{\partial \alpha_j}{\partial \theta_i}(\theta)\right)_{1\leq i,j\leq p} \quad \Longrightarrow \quad \operatorname{cov}_\theta(\widehat{T}) = \left(\frac{\partial \alpha_j}{\partial \theta_i}(\theta)\right)^{-1}_{1\leq i,j\leq p} \cdot I(\theta) \cdot {}^t\left(\frac{\partial \alpha_j}{\partial \theta_i}(\theta)\right)^{-1}_{1\leq i,j\leq p}$$

Par ailleurs, comme $\widehat{T}$ est un estimateur sans biais de $g(\theta)$ d'après la preuve de l'Inégalité de Cramer-Rao,

$$\mathbb{E}_\theta\left(\widehat{T}(.) \cdot {}^t\left(\frac{\partial}{\partial \theta}(\log L_\theta(.))\right)\right) = I_p$$

et en utilisant (3) que l'on multiplie par $\left(\frac{\partial}{\partial \theta}(\log L_\theta(.))\right)$, on obtient :

$$\mathbb{E}_\theta\left(\left(\frac{\partial}{\partial \theta}(\log L_\theta(.))\right) \cdot {}^t\left(\frac{\partial}{\partial \theta}(\log L_\theta(.))\right)\right) = \mathbb{E}_\theta\left(\left(\frac{\partial \alpha_j}{\partial \theta_i}(\theta)\right) \cdot \widehat{T} \cdot {}^t\left(\frac{\partial}{\partial \theta}(\log L_\theta(.))\right)\right) + \mathbb{E}_\theta\left(\frac{\partial \beta}{\partial \theta}(\theta) \cdot {}^t\left(\frac{\partial}{\partial \theta}(\log L_\theta(.))\right)\right),$$

et donc $I(\theta) = \left(\frac{\partial \alpha_j}{\partial \theta_i}(\theta)\right)_{1\leq i,j\leq p}$. A l'aide de cette égalité, et en reprenant le calcul précédent, on en arrive à ce que :

$$\operatorname{cov}_\theta(\widehat{T}) = I^{-1}(\theta),$$

donc $\widehat{T}$ atteint bien la borne de Cramer-Rao. De plus, grâce à (3),

$$\begin{aligned} \mathbb{E}_\theta\left(\frac{\partial}{\partial \theta}(\log L_\theta(x))\right) &= \mathbb{E}_\theta\left(\left(\frac{\partial \alpha_j}{\partial \theta_i}(\theta)\right)_{1\leq i,j\leq p} \cdot \widehat{T} + \frac{\partial \beta}{\partial \theta}(\theta)\right) \\ \text{soit} \qquad 0 &= \left(\frac{\partial \alpha_j}{\partial \theta_i}(\theta)\right)_{1\leq i,j\leq p} \theta + \frac{\partial \beta}{\partial \theta}(\theta) \\ \text{et donc} \qquad \theta &= -\left(\frac{\partial \alpha_j}{\partial \theta_i}(\theta)\right)^{-1}_{1\leq i,j\leq p} \cdot \frac{\partial \beta}{\partial \theta}(\theta). \end{aligned}$$

$\Longrightarrow$ D'après la preuve de l'Inégalité de Cramer-Rao, si $\widehat{T}$ est un estimateur sans biais de θ atteignant la borne de Cramer-Rao, alors

$$\operatorname{cov}_\theta(\widehat{T} - I^{-1}(\theta)\, Z_\theta) = 0.$$

Ainsi, pour tout $\theta \in \Theta$, il existe un ensemble $N_\theta \subset (\Omega')^n$ tel que $\mathbb{P}_\theta^{(n)}(N_\theta) = 1$ et tel que pour tout $x \in N_\theta$, $\widehat{T}(x) - \theta = I^{-1}(\theta)\, Z_\theta(x)$. Par le même procédé que celui de la preuve de la nullité de l'information de Fisher pour une statistique libre, on peut déterminer un ensemble N ne dépendant pas de θ, tel que cette propriété soit également vraie, avec $\mu(N) = 1$, ce qui revient à écrire que $\forall x \in N$,

$$I(\theta)\left(\widehat{T}(x) - \theta\right) = \frac{\partial}{\partial\theta}(\log L_\theta(x)), \quad \text{pour tout } \theta \in \Theta.$$

Alors en intégrant par rapport à θ, et en notant $\begin{cases} \alpha(\theta) \text{ le vecteur colonne "intégrant" } I(\theta)\cdot \\ \beta(\theta) \text{ la fonction "intégrant" } -I(\theta)\,\theta \\ b(x) \text{ une fonction ne dépendant pas de } \theta \end{cases}$ on a $\log L_\theta(x) = \alpha(\theta)\cdot\widehat{T}(x) + \beta(\theta) + b(x)$, d'où l'écriture de la vraisemblance sous forme d'un modèle exponentiel, et on retrouve l'expression de θ par le même raisonnement que plus haut. □

Corollaire. *A l'inverse, si l'on dispose d'un modèle exponentiel régulier (2), alors il n'existe qu'une seule fonction (à une transformation affine près) du paramètre pouvant être estimer efficacement, il s'agit de* $g(\theta) = -\frac{1}{n}\cdot\left(\frac{\partial\alpha_j}{\partial\theta_i}(\theta)\right)^{-1}_{1\le i,j\le p}\cdot\frac{\partial\beta}{\partial\theta}(\theta)$ *(noter que cette fonction semble dépendre de n; dans le cas de v.a.i.i.d. ce n'est pas le cas). L'estimateur est alors :* $\widehat{T} = \frac{1}{n}\cdot(a_1(X_1,\dots,X_n),\dots,a_p(X_1,\dots,X_n))$ *et sa matrice de covariance minimale est donnée par sa borne de Cramer-Rao, soit :*

$$cov_\theta(\widehat{T}) = \frac{1}{n}\cdot\frac{\partial g}{\partial\theta}(\theta)\cdot\left(\frac{\partial\alpha_j}{\partial\theta_i}(\theta)\right)^{-1}_{1\le i,j\le d}.$$

5.4 Estimateur du maximum de vraisemblance

Nous allons voir une méthode permettant d'obtenir aisément et dans la plupart des cas un estimateur possédant de très bonnes qualités... Par la suite on se place une nouvelle fois dans le cadre d'un modèle statistique paramétrique $((\Omega')^n, \mathcal{A}'_n, \mathbb{P}_\theta^{(n)}, \theta \in \Theta)$, avec $\Theta \subset \mathbb{R}^p$, dominé.

Définition. Pour $(x_1,\dots,x_n) \in (\Omega')^n$, soit $\theta \in \Theta \mapsto L_\theta(x_1,\dots,x_n)$ la vraisemblance du modèle. On appelle **estimateur du maximum de vraisemblance** une statistique $\widehat{\theta}_n$ telle que pour $(X_1,\dots,X_n)$ un n-échantillon quelconque du modèle:

$$L_{\widehat{\theta}_n}(X_1,\cdots,X_n) = \sup_{\theta\in\Theta} L_\theta(X_1,\cdots,X_n).$$

Remarque: Il n'y a pas de garantie de l'unicité d'un tel estimateur. Une méthode pour l'obtenir (mais pas toujours) est de rechercher un extremum local de L_θ sur Θ, ce qui pourra être fait en annulant les dérivées partielles de L_θ par θ_i. De même, il est clair que l'estimateur du maximum de vraisemblance pourra être également obtenu en maximisant le logarithme de la vraisemblance, appelé encore la log-vraisemblance.

Propriété. *On suppose que le modèle est régulier. Si on suppose qu'il existe un estimateur sans biais efficace de* θ *alors c'est l'estimateur du maximum de vraisemblance de* θ.

Proof. D'après ce qui précède, si le modèle est régulier et que $\widehat{T}$ est un estimateur sans biais efficace de θ, alors le modèle est exponentiel et l'égalité (3) a encore lieu, soit pour tout $\theta \in \Theta$,

$$\frac{\partial}{\partial\theta}(\log L_\theta(x)) = \left(\frac{\partial\alpha_j}{\partial\theta_i}(\theta)\right)_{1\le i,j\le p}\cdot\widehat{T} + \frac{\partial\beta}{\partial\theta}(\theta) \implies \left(\frac{\partial\alpha_j}{\partial\theta_i}(\theta)\right)_{1\le i,j\le p}\cdot\mathbb{E}_\theta(\widehat{T}) + \frac{\partial\beta}{\partial\theta}(\theta) = 0.$$

Comme $\widehat{T}$ est un estimateur sans biais de θ, on a donc $\left(\frac{\partial\alpha_j}{\partial\theta_i}(\theta)\right)_{1\le i,j\le p}\cdot\theta + \frac{\partial\beta}{\partial\theta}(\theta) = 0$, pour tout $\theta \in \Theta$, ce qui s'applique également à $\widehat{\theta}$ et donc :

$$\left(\frac{\partial\alpha_j}{\partial\theta_i}(\widehat{\theta})\right)_{1\le i,j\le p}\cdot\widehat{\theta} + \frac{\partial\beta}{\partial\theta}(\widehat{\theta}) = 0.$$

Mais d'après sa définition, le modèle étant régulier $\widehat{\theta}$ minimise la log-vraisemblance et annule donc sa dérivée, ce qui implique que :

$$\left(\frac{\partial\alpha_j}{\partial\theta_i}(\widehat{\theta})\right)_{1\le i,j\le p}\cdot\widehat{T} + \frac{\partial\beta}{\partial\theta}(\widehat{\theta}) = 0.$$

En conséquence, obtient :

$$\left(\frac{\partial \alpha_j}{\partial \theta_i}(\widehat{\theta})\right)_{1\leq i,j\leq p} \cdot \left(\widehat{T}-\widehat{\theta}\right) = 0 \implies \widehat{T}=\widehat{\theta},$$

car la matrice des dérivées des α_j est supposée de rang d. Enfin, l'unicité de $\widehat{\theta}$ est liée à l'écriture du modèle exponentiel. □

Nous allons nous intéresser maintenant au comportement asymptotique d'un estimateur du maximum de vraisemblance, donc quand la taille n de l'échantillon tend vers l'infini. Il est clair que pour chaque n l'expression de l'estimateur est différente et, surtout, le modèle statistique change. Pour palier à cela, on se placera dans un "gros" modèle, $((\Omega')^{\mathbb{N}}, \mathcal{A}'^{\mathbb{N}}, \mathbb{P}_\theta^{\mathbb{N}}, \theta\in\Theta)$, où $\Theta\subset\mathbb{R}^p$ (la dimension du paramètre reste constante) dans lequel un échantillon est une suite de v.a. Par ailleurs, on supposera désormais que **tout échantillon de ce modèle est constitué de v.a.i.i.d.**, et que $d\mathbb{P}_\theta^{\mathbb{N}} = (f_\theta\cdot d\mu)^{\otimes\mathbb{N}}$, le modèle étant dominé par la mesure μ, et f_θ étant la densité de chaque X_i par rapport à μ.

Théorème (Convergence de l'estimateur du maximum de vraisemblance)**.** *On suppose le modèle paramétrique $((\Omega')^{\mathbb{N}}, \mathcal{A}'^{\mathbb{N}}, (f_\theta\cdot d\mu)^{\otimes\mathbb{N}}, \theta\in\Theta)$, où $\Theta\subset\mathbb{R}^d$ est régulier. On suppose en plus que le modèle est identifiable (au sens où $f_{\theta_1}=f_{\theta_2}$, μ-presque partout, entraîne $\theta_1=\theta_2$). Alors si la suite $(X_n)_{n\in\mathbb{N}}$ est issue du modèle avec pour paramètre $\theta_0\subset\Theta$,*

$$\widehat{\theta}_n \xrightarrow[n\to\infty]{p.s.} \theta_0 \quad \textit{pour la mesure } (f_{\theta_0}\cdot d\mu)^{\otimes\mathbb{N}}.$$

Proof. En premier lieu, pour n fixé, il est clair que pour tout $\theta\in\Theta$:

$$\log(L_\theta(x_1,\dots,x_n)) - \log(L_{\theta_0}(x_1,\dots,x_n)) = \sum_{i=1}^n \log\left(\frac{f_\theta(x_i)}{f_{\theta_0}(x_i)}\right).$$

Par ailleurs, pour tout $i\in\mathbb{N}$, les X_i ont tous la même loi et pour $\theta\in\Theta$,

$$\begin{aligned}
\mathbb{E}_{\theta_0}\left[\log\left(\frac{f_\theta(X_i)}{f_{\theta_0}(X_i)}\right)\right] &\leq \log\left(\mathbb{E}_{\theta_0}\left[\frac{f_\theta(X_i)}{f_{\theta_0}(X_i)}\right]\right) \quad \text{(Inégalité de Jensen pour la fonction } -\log)\\
&\leq \log\left(\mathbb{E}_\mu\left[f_\theta(X_i)\right]\right)\\
&\leq 0.
\end{aligned}$$

En fait, du fait que la fonction $-\log$ est strictement convexe, la borne 0 ne peut être atteinte que si $f_\theta = f_{\theta_0}$. Ainsi, avec la contrainte d'un modèle identifiable, dès que $\theta\neq\theta_0$, alors :

$$\mathbb{E}_{\theta_0}\left[\log\left(\frac{f_\theta(X_i)}{f_{\theta_0}(X_i)}\right)\right] < 0.$$

On peut appliquer la loi forte des grands nombres pour les variables aléatoires $\left(\log\left(\frac{f_\theta(X_i)}{f_{\theta_0}(X_i)}\right)\right)_{i\in\mathbb{N}}$ (qui sont bien i.i.d. et $\mathbb{L}^1$ car le modèle est régulier), et ainsi:

$$\begin{aligned}
\frac{1}{n}\left(\log(L_\theta(X_1,\dots,X_n)) - \log(L_{\theta_0}(X_1,\dots,X_n))\right) &= \frac{1}{n}\sum_{i=1}^n \log\left(\frac{f_\theta(X_i)}{f_{\theta_0}(X_i)}\right)\\
&\xrightarrow[n\to\infty]{p.s.} \mathbb{E}_{\theta_0}\left[\log\left(\frac{f_\theta(X_i)}{f_{\theta_0}(X_i)}\right)\right] < 0,
\end{aligned}$$

la convergence presque sûre ayant lieu pour la mesure $(f_{\theta_0}\cdot d\mu)^{\otimes\mathbb{N}}$. Considérons maintenant pour tout $\varepsilon>0$ une famille dénombrable $(\theta_i^{(\varepsilon)})_{i\in I}$ dense sur la sphère de centre θ_0 et de rayon ε. Du fait du caractère dénombrable de cette famille, pour tout $\varepsilon>0$, il existe n_ε tel que pour tout $n\geq n_\varepsilon$, pour tout $i\in I$:

$$\log(L_{\theta_i^{(\varepsilon)}}(X_1,\dots,X_n)) < \log(L_{\theta_0}(X_1,\dots,X_n)) \quad \text{p.s. pour la mesure } (f_{\theta_0}\cdot d\mu)^{\otimes\mathbb{N}}.$$

Comme le modèle est régulier, pour tout $n\in\mathbb{N}^*$, la log-vraisemblance de $X_1,\dots,X_n$ est continue sur Θ. De plus pour tout n elle atteint son unique maximum en θ_0. En conséquence, pour $n\geq n_\varepsilon$, $\widehat{\theta}_n$ sera à l'intérieur de la boule de centre θ_0 et de rayon ε (toujours p.s. pour la mesure $(f_{\theta_0}\cdot d\mu)^{\otimes\mathbb{N}}$). Le raisonnement étant vrai pour tout $\varepsilon>0$, le théorème s'en déduit. □

Théorème (Normalité asymptotique de l'estimateur du maximum de vraisemblance)**.** *On suppose le modèle paramétrique* $((\Omega')^{\mathbb{N}}, \mathcal{A}'^{\mathbb{N}}, (f_\theta \cdot d\mu)^{\otimes \mathbb{N}}, \theta \in \Theta)$*, où* $\Theta \subset \mathbb{R}^p$ *est régulier. On suppose en plus que le modèle est identifiable et que la fonction* $\theta \in \Theta \mapsto L_\theta$ *est de classe* $\mathcal{C}^2(\Theta)$*. Alors si la suite* $(X_n)_{n\in\mathbb{N}}$ *est issue du modèle avec pour paramètre* $\theta_0 \subset \Theta$*:*

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}_d(0, I_1^{-1}(\theta_0)),$$

où $I_1(\theta)$ *est la matrice de Fisher de taille* p *(supposée inversible) pour la variable* X_1*.*

Proof. Comme le modèle est régulier, on peut différencier la vraisemblance et pour tout $\theta \in \Theta$, noter :

$$M_\theta(X_1, \ldots, X_n) = \frac{1}{n} \frac{\partial}{\partial \theta} \log L_\theta(X_1, \ldots, X_n) = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log (f_\theta(X_i)) .$$

Un développement limité d'ordre 1 de M_θ autour de θ_0 est possible (toujours en raison du modèle régulier) et donc pour tout tout $\theta \in \Theta$:

$$M_\theta(X_1, \ldots, X_n) = M_{\theta_0}(X_1, \ldots, X_n) + (\theta - \theta_0) \cdot \frac{\partial}{\partial \theta} M_{\theta^*}(X_1, \ldots, X_n),$$

avec θ^* dans le segment $[\theta, \theta_0]$ (remarquons que $\frac{\partial}{\partial \theta} M_{\theta^*}(X_1, \ldots, X_n)$ est une matrice carrée de taille d). Ainsi en remplaçant θ par $\widehat{\theta}_n$, on obtient pour chaque n l'existence de θ_n^* appartenant au segment $[\widehat{\theta}_n, \theta_0]$ tel que :

$$M_{\widehat{\theta}_n}(X_1, \ldots, X_n) = M_{\theta_0}(X_1, \ldots, X_n) + (\widehat{\theta}_n - \theta_0) \cdot \frac{\partial}{\partial \theta} M_{\theta_n^*}(X_1, \ldots, X_n). \tag{4}$$

Pour un modèle régulier, on a vu que $\mathbb{E}_{\theta_0}\left(\frac{\partial^2}{\partial \theta^2} \log f_{\theta_0}(X_i)\right) = -I_1(\theta_0)$, matrice de Fisher pour n'importe quelle variable X_i. Ainsi, $\frac{\partial}{\partial \theta} M_\theta(.)$ étant une moyenne empirique, on a par la loi forte des grands nombres:

$$\frac{\partial}{\partial \theta} M_{\theta_0}(X_1, \ldots, X_n) = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2}{\partial \theta^2} \log f_{\theta_0}(X_i) \xrightarrow[n\to\infty]{p.s.} -I_1(\theta_0) \text{ pour la mesure } (f_{\theta_0} \cdot d\mu)^{\otimes \mathbb{N}}.$$

Maintenant, en utilisant le fait que les densités f_θ sont de classe $\mathcal{C}^2(\Theta)$ et en utilisant la convergence presque sûre de $\widehat{\theta}_n$ vers θ_0 démontrée au théorème précédent, on a :

$$\frac{\partial}{\partial \theta} M_{\theta_n^*}(X_1, \ldots, X_n) \xrightarrow[n\to\infty]{p.s.} -I_1(\theta_0) \text{ pour la mesure } (f_{\theta_0} \cdot d\mu)^{\otimes \mathbb{N}}.$$

Finalement, comme $\widehat{\theta}_n$ est le maximum d'une fonction de classe $\mathcal{C}^1$, cet estimateur annule $M_{\widehat{\theta}_n}(X_1, \ldots, X_n)$, et donc l'égalité (4) devient :

$$M_{\theta_0}(X_1, \ldots, X_n) \cdot I_1^{-1}(\theta_0) = (\widehat{\theta}_n - \theta_0).$$

Enfin, comme $M_{\theta_0}(X_1, \ldots, X_n)$ est une moyenne empirique, ce vecteur aléatoire vérifie un théorème de la limite centrale :

$$\sqrt{n}\left(M_{\theta_0}(X_1, \ldots, X_n) - \mathbb{E}_{\theta_0}\left(\frac{\partial}{\partial \theta} \log f_{\theta_0}(X_i)\right)\right) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}_d(0, I_1(\theta_0)),$$

d'après la première définition de l'information de Fisher. Comme $\mathbb{E}_{\theta_0}\left(\frac{\partial}{\partial \theta} \log f_{\theta_0}(X_i)\right) = 0$ (voir les propriétés précédentes), on obtient la normalité asymptotique de $\widehat{\theta}_n$. □

Remarque: Sous ces hypothèses, l'estimateur du maximum de vraisemblance est asymptotiquement sans biais et efficace. Cependant, à n fixé, il peut avoir un biais et ne pas être un estimateur efficace.

5.5 Régions de confiance

En pratique, estimer un paramètre le plus souvent ne suffit pas. On aimerait connaître plus précisément quelle marge de sécurité on a sur la connaissance de ce paramètre.

Définition. On se place dans le cadre d'un modèle paramétrique $((\Omega')^n, \mathcal{A}'_n, \mathbb{P}_\theta, \theta \in \Theta)$, où $\Theta \subset \mathbb{R}^p$. Soit $\alpha \in]0,1[$ un nombre fixé a priori. On appelle région de confiance du paramètre θ de niveau $1-\alpha$ un sous-ensemble aléatoire $R_{1-\alpha}$ inclus dans $\mathbb{R}^p$ et défini sur $((\Omega')^n, \mathcal{A}'_n)$, tel que pour tout $\theta \in \Theta$, $\{(x_1, \ldots, x_n) \in (\Omega')^n, \theta \in R_{1-\alpha}(x_1, \ldots, x_n)\} \in \mathcal{A}'_n$ et :

$$\inf_{\theta \in \Theta} \left\{ \mathbb{P}_\theta^{(n)}(\theta \in R_{1-\alpha}) \right\} \geq 1 - \alpha. \tag{5}$$

Si un échantillon observé $(X_1(\omega), \ldots, X_n(\omega))$ est connu, $R_{1-\alpha}(X_1(\omega), \ldots, X_n(\omega))$ est appelé région de confiance observé. Dans le cas où le paramètre est un réel ($p = 1$), on pourra obtenir un intervalle de confiance.

Comment déterminer une région de confiance ? En premier lieu, il est clair que pour tout $\alpha \in]0,1[$, $R_{1-\alpha} \subset \Theta$ (en général, on choisit α proche de 0, et en particulier $\alpha = 0.05$ est très souvent utilisé). Une démarche possible pour la construction de région de confiance est la suivante : naturellement, on désirerait utiliser un estimateur $\widehat{T}$ convergent de θ, mais sa loi dépend en général de θ ce qui rend difficile (à part quelques exceptions) son utilisation directe. On préférera donc utiliser ce que l'on appelle une fonction pivotale $\pi(\widehat{T}, \theta)$, qui est une fonction mesurable d'un estimateur et de θ et qui est une statistique libre. On essayera alors d'écrire la propriété (5) sous la forme

$$\inf_{\theta \in \Theta} \left\{ \mathbb{P}_\theta^{(n)}(\pi(\widehat{T}, \theta) \in C_\alpha) \right\} \geq 1 - \alpha,$$

où C_α est une région déterministe. Aussi pourra-t-on ensuite construire la région de confiance en fonction des quantiles (souvent à $\alpha/2$ et $1-\alpha/2$) de la loi de la fonction pivotale. **Exemple:** Si le modèle est régulier, sous les conditions du théorème de normalité asymptotique du maximum de vraisemblance, on peut également montrer (théorème de Slutski) que

$$\pi(\widehat{\theta}_n, \theta_0) = \sqrt{n} \cdot (I_1(\widehat{\theta}_n))^{1/2} \cdot \left(\widehat{\theta}_n - \theta_0 \right) \xrightarrow[n \to \infty]{\mathcal{L}} \mathcal{N}_d(0, I_p),$$

où I_d est la matrice identité de taille p et $(I_1(\theta))^{1/2} \cdot (I_1(\theta))^{1/2} = I_1(\theta)$ pour tout $\theta \in \Theta$. Ainsi, si n est grand, on pourra assimiler la loi de $\pi(\widehat{\theta}_n, \theta_0)$ avec la loi normale centrée réduite multidimensionnelle. Or si $Z \sim \mathcal{N}_p(0, I_p)$, avec $q_{1-\alpha/2} > 0$ le quantile d'une loi normale centrée réduite réelle de niveau $1-\alpha/2$, tel que $P(Z \in [-q_{1-\alpha/2}, q_{1-\alpha/2}]^d) \geq 1-\alpha$. Aussi le polyèdre $n^{-1/2} \cdot (I_1(\widehat{\theta}_n))^{-1/2} \cdot [-q_{1-\alpha/2}, q_{1-\alpha/2}]^d$ recentré autour de $\widehat{\theta}_n$ formera la région de confiance cherchée.

6 Tests paramétriques

6.1 Principes d'un test paramétrique

Un test permet, à partir d'une réalisation d'un échantillon, de décider entre deux hypothèses, en mettant en avant une hypothèse privilégiée, appelée hypothèse H_0, et une hypothèse alternative, appelée H_1. On associe à un test un niveau α (avec souvent $\alpha \simeq 0.05$) et une puissance $1-\beta$. La plupart du temps, α est fixé a priori et β s'en déduit. Plus précisément,

Définition. On se place dans le cadre d'un modèle paramétrique dominé $((\Omega')^n, \mathcal{A}'_n, \mathbb{P}_\theta^{(n)}, \theta \in \Theta)$, où $\Theta \subset \mathbb{R}^p$ et soit θ la "vraie" valeur du paramètre. Un problème de test est un choix entre deux

hypothèses :

$$\begin{cases} H_0 : \theta \in \Theta_0 & : \text{ hypothèse dite nulle} \\ \\ H_1 : \theta \in \Theta_1 & : \text{ hypothèse dite alternative,} \end{cases} \tag{6}$$

où $\Theta_0 \subset \mathbb{R}^p$, $\Theta_1 \subset \mathbb{R}^d$ et $\Theta_0 \cap \Theta_1 = \emptyset$.

Ceci posé, on peut préciser deux types de problèmes de tests suivant les constitutions de Θ_0 et Θ_1 :

Définition. Une hypothèse (H_0 ou H_1) est dite simple si elle est associée à un singleton (Θ_0 ou Θ_1). Sinon, elle sera dite composite. Dans le cas réel ($\Theta \subset \mathbb{R}$), si H_0 est simple de la forme $\theta = \theta_0$, et si H_1 est composite de la forme $\theta > \theta_0$ ou $\theta < \theta_0$, on parlera de test unilatéral; si H_1 est composite de la forme $\theta \neq \theta_0$, on parlera de test bilatéral.

Comment faire pour choisir entre les deux hypothèses H_1 et H_2 ? Il faudra partir de ce que l'on peut connaître du modèle, c'est-à-dire généralement un échantillon observé $(X_1, \ldots, X_n)$. Pour cela, on définit une statistique qui sera la clé de voûte du test :

Définition. Dans le cadre du problème de test (6), soit $\widehat{T}$ une statistique (donc une fonction mesurable d'un échantillon $(X_1, \cdots, X_n)$ issu du modèle) à valeurs dans $\mathbb{R}^d$, qui sera appelée statistique du test. Le test sera défini par la fonction $\widehat{\phi} = \mathbb{I}_{\widehat{T} \in W}$, où W est une partie de $\mathbb{R}^p$ appelée région critique du test (et sa partie complémentaire dans $\mathbb{R}^p$ est appelée région d'acceptation du test). Si $\widehat{\phi} = 1$, on choisira H_1, sinon on décidera plutôt H_0.

Donc, à chaque hypothèse H_0 et H_1, on associe une partie de $\mathbb{R}^p$ pour la statistique de test $\widehat{T}$. En général, ces parties ne sont pas Θ_0 et Θ_1. Pour pouvoir précisément déterminer la région W, dans un cadre théorique (qui n'est pas le même que le cadre pratique, voir plus bas), on peut commencer par associer une fonction puissance à la statistique de test, puis définir les erreurs de premier espèce α et de deuxième espèce β :

Définition. Pour la statistique de test $\widehat{T}$, on associe :

- une fonction puissance, qui est la probabilité de choisir H_1 : $\theta \in \Theta_1 \mapsto \mathbb{P}_\theta^{(n)}(\widehat{T} \notin W)$.
- une erreur de première espèce : $\mathbb{P}_{H_0}(\text{Choisir } H_1) = \alpha = \sup_{\theta \in \Theta_0} \mathbb{P}_\theta^{(n)}(\widehat{T} \in W)$;
- une erreur de seconde espèce : $\mathbb{P}_{H_1}(\text{Choisir } H_0) = \beta = \sup_{\theta \in \Theta_1} \mathbb{P}_\theta^{(n)}(\widehat{T} \notin W)$.

La puissance du test est $1 - \beta$.

Cependant, ce qui vient d'être écrit reste théorique. En pratique, on utilisera plutôt la démarche suivante :

Construction concrète d'un test: On suppose le problème de test (6). On pose également a priori α qui dépend du problème posé (mais en général $\alpha = 0.05$), et $1 - \alpha$ est appelé le niveau du test. Par la suite, on réalise :

1. L'expression quantitative des hypothèses H_0 et H_1.
2. Le choix de la statistique $\widehat{T}$ du test.
3. La construction d'une région critique W à l'hypothèse H_1 par rapport à $\widehat{T}$.
4. La détermination explicite de W en fonction de α.

5. Le calcul (si possible) de la puissance du test $1-\beta$.

6. Pour la réalisation de l'échantillon, rejet ou acceptation de H_0.

Remarque: Cependant, en pratique on ne procède pas ainsi. On a donc deux types d'erreur. Le choix de l'hypothèse privilégiée est donc fondamental car le résultat d'un test n'est pas symétrique. Par exemple, supposons que l'on ait pour modèle $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), \mathcal{N}(\theta,1)^{\otimes n}, \theta \in \mathbb{R})$ et que l'on veuille tester $H_0 : \theta = 0$ contre $H_1 : \theta = 1$ à partir d'un échantillon $(X_1, \cdots, X_n)$ du modèle. Nous verrons pourquoi un peu plus loin, $\overline{X}_n$ est une statistique de test pertinente. Par exemple, si $n = 1$, et $X_1(\omega) = \overline{X}_1(\omega) = 0.8$, que va-t-on choisir entre H_0 et H_1 ? Naturellement, une région critique sera de la forme $[s, +\infty[$, où $s \in \mathbb{R}$, car $\overline{X}_n$ est un estimateur de θ. On détermine s à l'aide de α, puisque $\mathbb{P}_{H_0}(\text{Choisir } H_1) = \alpha = \mathbb{P}_0(\overline{X}_1 \geq s)$, donc par exemple, si $\alpha = 0.05$, $s \simeq 1.65$. Par suite, si $\overline{X}_1(\omega) = 0.8$, on accepte H_0 et l'erreur de seconde espèce est $\mathbb{P}_1(\overline{X}_1 < s) \simeq 0.74$, donc très élevée : le test n'est pas très discriminant. Maintenant, si on inverse H_0 et H_1, soit $H_0 : \theta = 1$ contre $H_1 : \theta = 0$, le même résultat $X_1(\omega) = 0.8$, conduit à accepter H_0, avec une erreur de second espèce encore $\simeq 0.74$. On obtient donc deux résultats opposés pour la même expérience aléatoire. Les hypothèses H_0 et H_1 ne sont clairement pas interchangeable.

La question qui se pose maintenant est de savoir comment trouver une statistique de test. Une idée naturelle dans ce cadre paramétrique serait d'utiliser un estimateur du maximum de vraisemblance.

6.2 Test de Wald

Un estimateur du maximum de vraisemblance permet d'associer à chaque hypothèse du test un ensemble de même "forme" que Θ_0 et Θ_1. Cependant, la difficulté est trouver la loi de l'estimateur du maximum de vraisemblance $\widehat{\theta}$ à n fixé. Si cela est possible, on utilisera directement $\widehat{\theta}$ comme statistique de test.

Sinon, de manière plus générale, on connaît la loi asymptotique de $\widehat{\theta}_n$ quand le modèle est régulier. Donc quand n est grand, on pourrait utiliser une loi normale comme approximation de la loi de $\widehat{\theta}_n$. Mais, un nouvel obstacle apparaît : la matrice de covariance asymptotique, qui est la matrice d'information de Fisher inverse, dépend du paramètre θ. Aussi va-t-on préférer utiliser la statistique de test $\widehat{T}$ suivante :

Définition. Soit un modèle paramétrique dominé **régulier** $((\Omega')^n, \mathcal{A}'_n, \mathbb{P}_\theta^{(n)}, \theta \in \Theta)$, où $\Theta \subset \mathbb{R}^p$. La **statistique de Wald** $\widehat{T}$ pour le test $H_0 : \theta = \theta_0$ contre $H_1 : \theta \in \Theta_1$ est:

$$\widehat{T}_n = n\, {}^t(\widehat{\theta}_n - \theta)\, I(\theta)\, (\widehat{\theta}_n - \theta).$$

Pour montrer "théoriquement" la pertinence de ce test, on va donc considérer la suite de tests $(\widehat{T}_n)$ en se plaçant dans le "grand" modèle asymptotique :

Théorème. *Dans le cadre d'un modèle paramétrique régulier* $((\Omega')^{\mathbb{N}}, \mathcal{A}'_{\mathbb{N}}, (f_\theta \cdot d\mu)^{\otimes \mathbb{N}}, \theta \in \Theta)$*, où* $\Theta \subset \mathbb{R}^p$*, pour le problème de test* $H_0 : \theta = \theta_0$ *contre* $H_1 : \theta \neq \theta_0$*, alors, en notant* $\widehat{T}_n$ *la statistique de test de Wald pour le modèle projeté de taille* n *sous l'hypothèse* H_0*,*

$$\widehat{T}_n \xrightarrow[n\to\infty]{\mathcal{L}} \chi^2(p).$$

La région de rejet asymptotique du test sera donc de la forme $\widehat{T}_n > s_\alpha$*, où* s_α *est le quantile d'ordre* $1-\alpha$ *de la loi du* $\chi^2(p)$*: la suite de test* $(\widehat{T}_n)_n$ *a une puissance qui tend vers* 1 *lorsque* α *est fixé.*

Proof. La loi asymptotique de $\widehat{\theta}_n$ induit la loi asymptotique de $\widehat{T}_n$, car $\sqrt{n} \cdot I(\theta)^{1/2} \cdot (\widehat{\theta}_n - \theta)$ suit asymptotiquement une loi $\mathcal{N}(0, I_d)$ sous l'hypothèse H_0 et $\widehat{T}_n = \|\sqrt{n} \cdot I(\theta)^{1/2} \cdot (\widehat{\theta}_n - \theta)\|^2$. □

Voici donc un premier type de test, qui sous certaines conditions de régularités du modèle et pour certaines hypothèses de tests est intéressant. Mais pourrait-on faire mieux ? Et en quel sens ? Désormais, il nous faut donc définir un moyen de comparaison entre deux tests.

6.3 Test du rapport de vraisemblance

Définition. Sous les hypothèses précédentes, si $L_\theta(.)$ est la vraisemblance, on appellera test du rapport de vraisemblance (test de Neyman-Person dans le cas d'hypothèses simples) un test de statistique $\widehat{T}$ telle que:

$$\widehat{T} = \frac{\sup_{\theta \in \Theta_0} L_\theta(X_1, \cdots, X_n)}{\sup_{\theta \in \Theta_1} L_\theta(X_1, \cdots, X_n)}.$$

La région critique W associée à un tel test est de la forme $W =]-\infty, K[$ (donc si $\widehat{T} < K$, on rejette H_0).

Une des vertus du test du rapport de vraisemblance par rapport au test de Wald est qu'il peut être utilisé dans un modèle non régulier (mais la question de sa loi, ou de la loi d'une fonctionnelle de ce test, demeure).

Cependant, un tel test pour un modèle régulier, va pouvoir être traité de manière générale grâce à la normalité asymptotique de l'estimateur du maximum de vraisemblance:

Théorème. *Dans le cadre d'un modèle paramétrique régulier* $((\Omega')^{\mathbb{N}}, \mathcal{A}'^{\mathbb{N}}, (f_\theta \cdot d\mu)^{\otimes \mathbb{N}}, \theta \in \Theta)$*, où* $\Theta \subset \mathbb{R}^p$*, pour le problème de test* $H_0 : \theta = \theta_0$ *contre* $H_1 : \theta \neq \theta_0$*, alors, en notant* $\widehat{T}_n$ *la statistique du rapport de vraisemblance pour le modèle projeté de taille* n*,*

$$-2\log(\widehat{T}_n) \xrightarrow[n\to\infty]{\mathcal{L}} \chi^2(p).$$

La région de rejet asymptotique du test sera donc de la forme $-2\log(\widehat{T}_n) > s_\alpha$*, où* s_α *est le quantile d'ordre* $1-\alpha$ *de la loi du* $\chi^2(p)$*. La suite de test* $(\widehat{T}_n)_n$ *a donc une puissance qui tend vers* 1 *lorsque* α *est fixé.*

Proof. La démonstration reprend un peu celle de la normalité asymptotique du maximum de vraisemblance. □