UFR IM²AG

UNIVERSITÉ Grenoble Alpes

Polycopié pour l'UE Statistique.

Vincent Brault

Introduction

Ce cours repose principalement sur des notes de cours de Gérard Biau (Université Pierre et Marie Curie), Anatoli Juditsky (Université Grenoble Alpes), Jean-François Le Gall (Université Paris-Saclay) et Marie-Anne Poursat (Université Paris-Saclay).

Par défaut, les corrections des exercices ne sont pas fournies : soit votre raisonnement est juste du début jusqu'à la fin et, dans ce cas, votre résultat est bon ; soit vous hésitez sur l'une des justifications et même si le résultat est bon, il ne vaut rien. En revanche, lorsque des astuces existent, elles sont mises en fin de chapitre.

Enfin, merci aux étudiants qui me font des retours sur mes coquilles ou fautes d'inattention et, en particulier, à Ludovic Mailley, Romain Simeon et Maud Tartry.

Planning

Le planning mis ici est à titre indicatif et sera régulièrement mis à jour en fonction de l'avancée. Les dates peuvent également évoluer.

Cours	Date	Estimation
1	7-09-2018	Cours inversé pour réactiver les connaissances en probabilités.
2	14-09-2018	Cours/TD sur les vecteurs gaussiens.
3	21-09-2018	Cours/TD sur la fin des vecteurs gaussiens et sur les concepts fondamentaux de la Statistique.
4	28-09-2018	Cours/TD sur l'estimation : consistance d'un estimateur, estimateur des moments. *Distribution du DM.*
5	5-10-2018	Cours/TD sur l'estimation : estimateur du maximum de vraisemblance, qualité d'un estimateur
6	12-10-2018	TP sur machine sur la simulation de loi et l'estimation.
7	19-10-2018	Cours/TD sur l'estimation : Information de Fisher et borne de Cramer-Rao
8	26-10-2018	**Partiel**.
9	9-11-2018	Cours/TD sur les régions de confiance
10	16-11-2018	TP sur machine sur les régions de confiance et les tests d'hypothèses.
11	23-11-2018	Cours/TD sur les tests d'hypothèses

Table des matières

Chapitre 1

Rappels élémentaires de théorie des probabilités

"Quels que soient les progrès des connaissances humaines, il y aura toujours place pour l'ignorance et par suite pour le hasard et la probabilité."

Émile Borel

1.1 Objectifs

L'objectif de ce chapitre est de reprendre quelques bases nécessaires à la compréhension de la statistique et de l'UE Processus stochastiques.

1.2 Rappels/Pré-requis

Dans ce chapitre, nous aurons besoin des notions suivantes.

Analyse : opérations classiques sur les intégrales (notamment l'intégration par partie et les changements de variables), théorie sur les séries entières. Même si je fais un rappel sur la théorie de la mesure, je suppose que certains théorèmes sont connus (comme le théorème de convergence monotone largement utilisé dans les démonstrations) [1].

Algèbre : espace matriciel.

1. Pour celles et ceux qui auraient besoin de rappels plus conséquents, je recommande le polycopié excessivement complet de Le Gall (2006) disponible à cette adresse : `https://www.math.u-psud.fr/~jflegall/IPPA2.pdf`

1.3 Rappels généraux

Dans cette partie, nous commençons par rappeler le vocabulaire utilisé.

1.3.1 Théorie de la mesure

Commençons par introduire la notion de tribus nécessaire pour celle de probabilité.

Définitions 1 (Tribu et ensemble probabilisable)**.**

Soit Ω un ensemble quelconque, un ensemble $\mathcal{F}$ de parties de Ω est une **tribu** ou **σ-algèbre** si nous avons les conditions suivantes :

- $\mathcal{F}$ est non vide.
- $\mathcal{F}$ est stable par complémentaire :

$$\forall F \in \mathcal{F},\ F^C \in \mathcal{F}.$$

- $\mathcal{F}$ est stable par union dénombrable :

$$\forall (F_n)_{n\in\mathbb{N}} \in \mathcal{F}^{\mathbb{N}},\ \left(\bigcup_{n\in\mathbb{N}} F_n\right) \in \mathcal{F}.$$

Le couple $(\Omega, \mathcal{F})$ est un ensemble dit **probabilisable** ou **mesurable**.

Exemple

La **tribu borélienne** d'un ensemble Ω est la plus petite tribu contenant tous les ouverts de Ω. En particulier, lorsque $\Omega = \mathbb{R}$, la tribu borélienne est engendrée par tous les intervalles de $\mathbb{R}$.

Une mesure de probabilité est une mesure particulière.

Définitions 2 (Mesure et probabilité)**.**

Étant donné un ensemble probabilisable $(\Omega, \mathcal{F})$, une application $\mu : \mathcal{F} \to [0, +\infty]$ est une **mesure** si :

- $\mu(\emptyset) = 0$ où $\emptyset$ est l'ensemble vide.
- Soit $(F_n)_{n\in\mathbb{N}}$ une famille dénombrable d'ensembles de $\mathcal{F}$ deux à deux disjoints, c'est-à-dire que pour tout $n \neq m$, $F_n \cap F_m = \emptyset$, alors nous avons :

$$\mu\left(\bigsqcup_{n\in\mathbb{N}} F_n\right) = \sum_{n\in\mathbb{N}} \mu(F_n).$$

  Cette propriété est la σ**-additivité**.

Nous disons que la mesure $\mu : \mathcal{F} \to [0, +\infty[$ est **finie** si $\mu(\Omega) < +\infty$.
Une mesure finie $\mu : \mathcal{F} \to [0, 1]$ est une **mesure de probabilité** si $\mu(\Omega) = 1$.
Enfin, un triplet $(\Omega, \mathcal{F}, \mu)$ est un **espace mesuré**.

La proposition suivante permet de ne pas chercher à avoir forcément que la masse de l'ensemble global $\mu(\Omega)$ vaille forcément 1.

Astuce (Lien entre une mesure finie et une probabilité)**.**

Étant donné un ensemble probabilisable $(\Omega, \mathcal{F})$, si une application μ est une mesure finie non nulle alors l'application $\nu(\cdot) = \frac{\mu(\cdot)}{\mu(\Omega)}$ est une mesure de probabilité.

Preuve

Comme μ est une mesure, ν est également une mesure. Comme $\mu(\Omega) \in]0; +\infty[$, nous avons :

$$\nu(\Omega) = \frac{\mu(\Omega)}{\mu(\Omega)} = 1.$$

Définition 3 (Espace de probabilité)**.**

Soit Ω un ensemble quelconque, $\mathcal{F}$ une tribu et $\mathbb{P}$ une mesure de probabilité sur $\mathcal{F}$ alors $(\Omega, \mathcal{F}, \mathbb{P})$ est appelé **espace de probabilité**.

Avant de rappeler ce qu'est une variable aléatoire, nous avons besoin de la définition d'une fonction mesurable.

Définition 4 (Fonction mesurable)**.**

Étant donnés deux espaces probabilisables $(E, \mathcal{E})$ et $(F, \mathcal{F})$, la fonction $f : (E, \mathcal{E}) \to (F, \mathcal{F})$ est dite **mesurable** si l'image réciproque de $\mathcal{F}$ par f est incluse dans $\mathcal{E}$:

$$\forall B \in \mathcal{F},\ f^{-1}(B) \in \mathcal{E}.$$

Proposition 1 (Mesure image)**.**

Étant donnés deux espaces probabilisables $(E, \mathcal{E})$ et $(F, \mathcal{F})$ et une fonction $f : (E, \mathcal{E}) \to (F, \mathcal{F})$ mesurable, s'il existe une mesure μ sur l'espace $(E, \mathcal{E})$ alors l'application ν définie sur $\mathcal{F}$ par

$$\forall B \in \mathcal{F},\ \nu(B) = \mu\left(f^{-1}(B)\right)$$

est une mesure appelée **mesure image**.

Exercices 1.1

Si μ est une mesure finie, montrez que ν est finie également.

Pour clôturer cette section sur la théorie de la mesure, nous rappelons le théorème de Borel-Cantelli ; pour cela nous avons besoin de rappeler la définition suivante.

Définition 5 (Limite supérieure d'une suite de parties)**.**

Étant donné un ensemble Ω, la **limite supérieure** de la suite $(A_n)_{n\in\mathbb{N}}$ de parties de Ω notée $\limsup_{n\to+\infty} A_n$ est l'ensemble défini par :

$$\limsup_{n\to+\infty} A_n = \{\omega \in \Omega \,|\exists k_1 < k_2 < \cdots < k_\ell < \cdots \text{ tel que } \forall \ell \in \mathbb{N}, \omega \in A_{k_\ell}\}.$$

Une autre façon de le voir est d'utiliser les intersections et unions d'ensembles :

$$\limsup_{n\to+\infty} A_n = \bigcap_{n\in\mathbb{N}} \left(\bigcup_{k\geq n} A_k \right).$$

Remarque

Autrement dit, un élément ω appartient à la limite supérieure d'une suite de parties s'il appartient au moins à une infinité de ces parties car, dans ce cas, pour tout $k \in \mathbb{N}$, il existera toujours un indice ℓ plus grand que k tel que ω soit dans A_ℓ.

Théorème 2 (Borel-Cantelli)**.**

Étant donné $(\Omega, \mathcal{F}, \mu)$ un espace mesuré et $(A_n)_{n\in\mathbb{N}}$ une suite d'éléments de $\mathcal{F}$, si nous avons

$$\sum_{n\in\mathbb{N}} \mu(A_n) < +\infty$$

c'est-à-dire que la somme des mesures des parties de la suite est finie alors nous avons :

$$\mu\left(\limsup_{n\to+\infty} A_n\right) = 0.$$

Une autre façon de le voir est de dire que si la somme des probabilités d'une suite d'événements $(A_n)_{\in\mathbb{N}}$ est finie alors la probabilité qu'une infinité d'entre eux se réalise simultanément est nulle.

Preuve

Commençons par introduire pour tout $n \in \mathbb{N}$ l'ensemble des unions

d'ensembles indexés à partir de n :

$$B_n = \bigcup_{k \geq n} A_k.$$

Cette suite d'ensembles est mesurable par définition d'une tribu et décroissante car $B_n = A_n \cup B_{n+1}$ pour tout $n \in \mathbb{N}$. Comme la somme des $\mu(A_n)$ est finie, la mesure μ est finie sur l'union des ensembles A_n à laquelle appartient les ensembles B_n et nous avons donc :

$$\mu\left(\bigcap_{n \in \mathbb{N}} B_n\right) = \lim_{n \to +\infty} \mu(B_n).$$

De plus, nous avons pour tout $n \in \mathbb{N}$:

$$\begin{aligned} \mu(B_n) &= \mu\left(\bigcup_{k \geq n} A_k\right) \\ &\leq \sum_{k \geq n} \mu(A_k) \end{aligned}$$

qui est le reste de la série convergente de suite $\mu(A_n)$ donc qui tend vers 0 lorsque n tend vers l'infini.

En conclusion, la limite des $\mu(B_n)$ vaut 0 et, par définition, nous avons :

$$\begin{aligned} \mu\left(\limsup_{n \to +\infty} A_n\right) &= \mu\left[\bigcap_{n \in \mathbb{N}} \left(\bigcup_{k \geq n} A_k\right)\right] \\ &= \mu\left(\bigcap_{n \in \mathbb{N}} B_n\right) \\ &= \lim_{n \to +\infty} \mu(B_n) \\ &= 0. \end{aligned}$$

1.3.2 Variables aléatoires

Ceci est souvent oublié mais une variable aléatoire est en fait une fonction mesurable particulière.

Définition 6 (Variable aléatoire)**.**

Étant donné $(\Omega, \mathcal{F}, \mathbb{P})$ un espace de probabilité où Ω est un ensemble quelconque, $\mathcal{F}$ une tribu de Ω et $\mathbb{P}$ une mesure de probabilité sur $\mathcal{F}$, une **variable aléatoire** X est une fonction mesurable $X : (\Omega, \mathcal{F}) \to (E, \mathcal{E})$. On dira qu'elle est **réelle** si l'espace d'arrivé est $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ où $\mathcal{B}(\mathbb{R})$ est la tribu borélienne sur $\mathbb{R}$.

⚠ Attention au piège

Lorsque nous parlons de probabilité d'un événement, nous notons souvent et par exemple $\mathbb{P}(X = x)$ ou $\mathbb{P}(a \leq X \leq b)$. Ceci est un abus de notations symbolisant :

$$\begin{aligned} \mathbb{P}(X = x) &= \mathbb{P}\left(X^{-1}(\{x\})\right) \\ &= \mathbb{P}(\{\omega \in \Omega \,|\, X(\omega) = x\}) \end{aligned}$$

où $X^{-1}(\{x\})$ est l'image réciproque de l'ensemble singleton $\{x\}$. La mesure $\mathbb{P}$ existe donc sur Ω mais c'est la mesure image que nous utilisons.

Définition 7 (Fonction de répartition)**.**

Étant donnés l'espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$ et X une variable aléatoire à valeur dans $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, la **fonction de répartition** associée $F : \mathbb{R} \to [0, 1]$ est définie pour tout $x \in \mathbb{R}$ par $F(x) = \mathbb{P}(X \leq x)$.

Remarques

La fonction de répartition est parfois notée F_X pour préciser à quelle loi elle se rapporte lorsque plusieurs lois sont en présence.

Propriétés 3 (Propriétés de la fonction de répartition)**.**

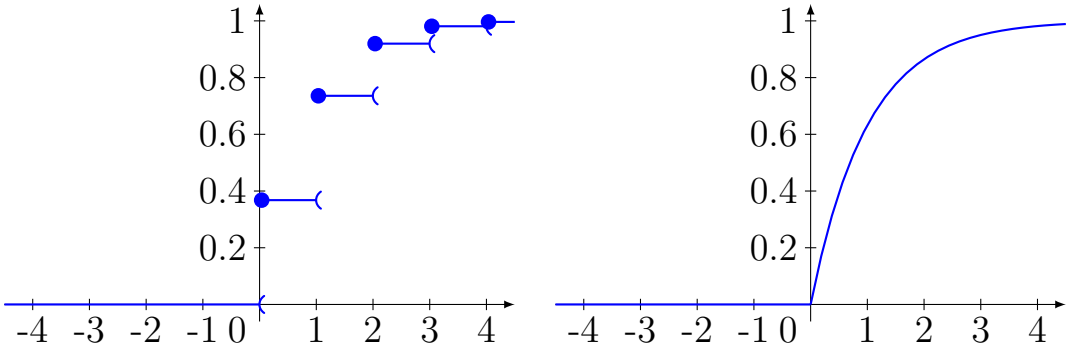
La fonction de répartition est croissante, finie à droite, continue à gauche et possède les limites suivantes :

$$\lim_{x \to -\infty} F(x) = 0 \text{ et } \lim_{x \to +\infty} F(x) = 1.$$

Exemple

Les fonctions de répartition des lois de Poisson $\mathcal{P}(1)$ et exponentielle $\mathcal{E}(1)$ ont les représentations suivantes :

Loi de Poisson $\mathcal{P}(1)$ Loi exponentielle $\mathcal{E}(1)$



Définitions 8 (Variables discrètes et variables continues)**.**

Nous distinguons deux grandes familles de variables aléatoires réelles :

- Les **variables discrètes** prennent leurs valeurs dans un ensemble fini ou dénombrable S. Dans ce cas, la loi est entièrement définie par la probabilité de chaque élément de l'ensemble S.
- Les **variables continues** admettent une densité par rapport à la mesure de Lebesgue sur $\mathbb{R}$, c'est-à-dire une fonction positive ou nulle, intégrable et d'aire sous la courbe valant 1. Dans ce cas, la fonction de répartition F est dérivable presque partout sur $\mathbb{R}$ et sa dérivée

$$\begin{aligned} f : \mathbb{R} &\to \mathbb{R}^+ \\ x &\mapsto F'(x) \end{aligned}$$

est appelée **la densité de probabilité**.

Remarque

Nous parlons de loi continue car la fonction de répartition est continue contrairement au cas des lois discrètes où elles sont constantes par morceaux.

Exemple

Exemples de lois discrètes :

- La loi de **Bernoulli** $\mathcal{B}(p)$ de paramètres $p \in]0;1[$ définie sur $\{0,1\}$ par $\mathbb{P}(X=1)=p$ et $\mathbb{P}(X=0)=1-p$.

- La loi de **Poisson** $\mathcal{P}(\lambda)$ de paramètre $\lambda > 0$ définie pour tout $k \in \mathbb{N}$ par :
$$\mathbb{P}(X=k) = e^{-\lambda}\frac{\lambda^k}{k!}$$
où $k!$ est le factoriel de k c'est-à-dire le produit de tous les entiers allant de 1 à k (avec la convention $0!=1$).

- La loi **uniforme** $\mathcal{U}(S)$ sur un ensemble fini S définie pour tout $x \in S$ par
$$\mathbb{P}(X=x) = \frac{1}{|S|}$$
où $|S|$ est le cardinal de l'ensemble S (c'est-à-dire son nombre d'éléments) est une loi discrète.

Exemples de lois continues :

- La loi **normale** ou loi **gaussienne** $\mathcal{N}(\mu,\sigma^2)$ de paramètres $(\mu,\sigma) \in \mathbb{R}\times\mathbb{R}^+$ de densité
$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}.$$

- La loi **exponentielle** $\mathcal{E}(\lambda)$ de paramètre $\lambda > 0$ de densité
$$f(x) = \lambda e^{-\lambda x}\mathbb{1}_{\mathbb{R}^+}(x) = \begin{cases} \lambda e^{-\lambda x} & \text{si } x \geq 0, \\ 0 & \text{sinon,} \end{cases}$$
où $\mathbb{1}_A(\cdot)$ est l'**indicatrice** de l'ensemble $A \subset \mathbb{R}$ valant 1 sur l'ensemble A et 0 sinon c'est-à-dire que pour tout $x \in \mathbb{R}$, nous avons
$$\mathbb{1}_A(x) = \begin{cases} 1 & \text{si } x \in A, \\ 0 & \text{sinon.} \end{cases}$$

- La loi **uniforme** $\mathcal{U}([a;b])$ sur un intervalle borné $[a;b]$ avec $a<b$ de densité
$$f(x) = \begin{cases} \frac{1}{b-a} & \text{si } x \in [a;b], \\ 0 & \text{sinon.} \end{cases}$$

Contre-exemple

Plusieurs densités peuvent mener à la même loi. Par exemple, la densité définie sur $\mathbb{R}$ par

$$f(x) = \begin{cases} |x| & \text{si } x \in \mathbb{Q}, \\ \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} & \text{sinon,} \end{cases}$$

fournit une variable aléatoire de même loi qu'une loi normale $\mathcal{N}(\mu, \sigma^2)$ puisque l'ensemble $\mathbb{Q}$ est de mesure de Lebesgue nulle. Notons au passage que rien n'impose à une densité d'être majorée ou même continue presque partout.

Ce cas étant pathologique, nous utiliserons dans ce cours essentiellement des densités de probabilité.

Point méthode.

Il est important de toujours vérifier que toutes les probabilités sont positives et que la somme sur l'espace de probabilité vaut 1.

⚠ Attention au piège

Bien que nous voyons généralement des densités discrètes ou continues en cours, ce ne sont pas les seules qui existent. Par exemple, la loi de probabilité du temps d'attente avant de passer à une caisse de supermarché peut être décrite de la façon suivante :

- La caisse est vide avec une probabilité $p \in]0; 1[$.
- S'il y a déjà des personnes, le temps d'attente suit une loi exponentielle $\mathcal{E}(\lambda)$ de paramètre $\lambda > 0$.

Nous voyons que la loi est définie sur $\mathbb{R}^+$ (donc infini non-dénombrable) mais la valeur 0 possède une masse p et il n'existe donc pas de densité par rapport à la mesure de Lebesgue associée à cette loi.

1.4 Caractéristique des variables aléatoires

Dans cette partie, nous rappelons ce qui caractérisent les variables aléatoires.

1.4.1 Moments des variables aléatoires

Définition 9 (Espérance mathématique d'une variable aléatoire)**.**

L'**espérance mathématique** de la variable aléatoire X est définie

- dans le cas discret sur l'ensemble fini ou dénombrable S par :

$$\mathbb{E}[X] = \sum_{x \in S} x \mathbb{P}(X = x).$$

- dans le cas continu sur l'ensemble non-dénombrable S par :

$$\mathbb{E}[X] = \int_S x f(x) \mathrm{d}x.$$

Remarques

L'espérance mathématique est parfois notée $\mathbb{E}_X[\cdot]$ pour préciser sous quelle loi elle est calculée lorsque plusieurs lois sont en présence.

Définitions 10 (Moment des variables aléatoires)**.**

Plus généralement, pour toute fonction $h : S \to S'$ et toute variable aléatoire X à valeurs dans $\mathbb{R}$, nous définissons l'espérance de $h(X)$ par :

- dans le cas discret sur l'ensemble fini ou dénombrable S par :

$$\mathbb{E}\left[h(X)\right] = \sum_{x \in S} h(x)\mathbb{P}\left(X = x\right).$$

- dans le cas continu sur l'ensemble non-dénombrable S par :

$$\mathbb{E}\left[h(X)\right] = \int_S h(x)f(x)\mathrm{d}x.$$

En particulier, pour tout $k \in \mathbb{N}$, nous appelons

- **moment d'ordre** k l'espérance $\mathbb{E}\left[X^k\right]$,
- **moment centré d'ordre** k l'espérance $\mathbb{E}\left[\left(X - \mathbb{E}\left[X\right]\right)^k\right]$,
- **moment absolu d'ordre** k l'espérance $\mathbb{E}\left[\left|X\right|^k\right]$,
- **moment absolu centré d'ordre** k l'espérance $\mathbb{E}\left[\left|X - \mathbb{E}\left[X\right]\right|^k\right]$.

Ces moments sont définis sous condition que les espérances soient finies.

⚠ Attention au piège

Les moments centrés d'ordre $k \geq 1$ ne sont définis qu'à condition que la variable aléatoire ait une espérance finie.

Exercices 1.2

Calculer les moments d'ordre k de la loi exponentielle $\mathcal{E}(\lambda)$.
Pour toute variable aléatoire réelle continue de densité symétrique par rapport à l'origine, c'est-à-dire avec une densité paire, montrez que les moments d'ordre impair sont nuls lorsqu'ils sont définis. Donner un exemple de variable aléatoire réelle continue de densité symétrique dont les moments d'ordre impair ne sont pas définis.

Proposition 4.

Soit une variable aléatoire X réelle possédant un moment d'ordre ℓ alors pour tout $0 \leq k \leq \ell$, elle possède un moment d'ordre k.

Preuve

Il s'agit de décomposer $\mathbb{R}$ entre l'intervalle $[-1;1]$ où x^ℓ sera plus grand en valeur absolu que x^k et son complémentaire. Pour qu'il y ait un moment, il faut que les intégrale sur $]-\infty;0]$ et sur $[0;+\infty[$ soient finies. Nous allons présenter la démonstration sur $[0;+\infty[$, celle sur $]-\infty;0]$ sera identique au signe près suivant la parité de k pour le cas de variable continue :

$$\begin{aligned}
\int_0^{+\infty} x^k f(x)\mathrm{d}x &= \int_0^1 \underbrace{x^k}_{\leq 1} f(x)\mathrm{d}x + \int_1^{+\infty} \underbrace{x^k}_{\leq x^\ell} f(x)\mathrm{d}x \\
&\leq \int_0^1 1 f(x)\mathrm{d}x + \int_1^{+\infty} x^\ell f(x)\mathrm{d}x \\
&\leq 1 + \int_1^{+\infty} x^\ell f(x)\mathrm{d}x \\
&< +\infty.
\end{aligned}$$

Définitions 11 (Moyenne et variance)**.**

Étant donnée une variable aléatoire X, nous appelons **moyenne** son moment d'ordre un $\mathbb{E}[X]$, **variance**, notée $\mathbb{V}[X]$ ou $Var(X)$, son moment centré d'ordre deux $\mathbb{E}\left[(X - \mathbb{E}[X])^2\right]$ et **écart-type**, noté σ_X, la racine de la variance.

Proposition 5.

La variance peut aussi s'écrire comme la différence entre le moment d'ordre 2 et le carré du moment d'ordre 1 :

$$\mathbb{V}[X] = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2.$$

Preuve

Ceci vient de la linéarité de l'espérance et du fait que l'espérance

est un scalaire :

$$\begin{aligned}\mathbb{V}[X] &= \mathbb{E}\left[(X-\mathbb{E}[X])^2\right] \\ &= \mathbb{E}\left[X^2-2X\mathbb{E}[X]+\mathbb{E}[X]^2\right] \\ &= \mathbb{E}\left[X^2\right]-\mathbb{E}\left[2X\mathbb{E}[X]\right]+\mathbb{E}\left[\mathbb{E}[X]^2\right] \\ &= \mathbb{E}\left[X^2\right]-2\mathbb{E}[X]\mathbb{E}[X]+\mathbb{E}[X]^2 \\ &= \mathbb{E}\left[X^2\right]-\mathbb{E}[X]^2.\end{aligned}$$

1.4.2 Indépendance de variables aléatoires

Cette partie est construite pour donner les outils pour démontrer la *loi du zéro-un de Borel*.

Définition 12 (Indépendance de n variables aléatoires)**.**

Étant donné $(\Omega,\mathcal{F},\mathbb{P})$ un espace de probabilité et n variables aléatoires $(X_1,\ldots,X_n)$ à valeurs dans les espaces mesurés $(E_i,\mathcal{E}_i)_{i\in\{1,\ldots,n\}}$, nous disons que les variables sont **indépendantes** si :

$$\forall(A_1,\ldots,A_n)\in\mathcal{E}_1\times\cdots\times\mathcal{E}_n,\ \mathbb{P}(X_1\in A_1,\ldots,X_n\in A_n)=\prod_{i=1}^{n}\mathbb{P}(X_i\in A_i).$$

Ou, de façon équivalente, si pour toute famille de fonctions $(g_i)_{i\in\{1,\ldots,n\}}$ telle que pour chaque $i\in\{1,\ldots,n\}$, g_i est mesurable allant de E_i vers $\mathbb{R}$ alors les variables aléatoires doivent vérifier :

$$\mathbb{E}\left[\prod_{i=1}^{n}g_i(X_i)\right]=\prod_{i=1}^{n}\mathbb{E}[g_i(X_i)]$$

à condition que les espérances aient un sens.

Remarque

Le lecteur curieux de connaître une démonstration de l'équivalence peut se reporter au théorème 9.2.1 du polycopié de Le Gall (2006) disponible à l'adresse suivante : `https://www.math.u-psud.fr/~jflegall/IPPA2.pdf`.

En particulier, nous avons l'habitude de faire les démonstrations suivantes pour les lois discrètes et continues.

Point méthode.

Si X et Y sont deux variables aléatoires réelles, nous dirons que X et Y sont **indépendantes** si et seulement si pour toutes fonctions boréliennes g et h, nous ayons :

$$\mathbb{E}\left[g(X)h(Y)\right] = \mathbb{E}\left[g(X)\right]\mathbb{E}\left[h(Y)\right]$$

ou encore si et seulement si pour tout $(x, y) \in \mathbb{R}^2$, nous avons :

$$\mathbb{P}\left(X \leq x, Y \leq y\right) = \mathbb{P}\left(X \leq x\right)\mathbb{P}\left(Y \leq y\right).$$

En particulier, si $(X_1, \ldots, X_n)$ sont n variables aléatoires *discrètes* à valeurs dans des ensembles finis ou dénombrables $S_1, \ldots, S_n$, nous dirons que les variables sont **indépendantes** si et seulement si nous avons :

$$\forall (x_1, \ldots, x_n) \in \prod_{i=1}^{n} S_i,\ \mathbb{P}\left(X_1 = x_1, \ldots, X_n = x_n\right) = \prod_{i=1}^{n} \mathbb{P}\left(X_i = x_i\right).$$

Astuce.

Il est également possible de tester facilement l'indépendance entre n'importe quelle variable discrète X à valeurs dans un ensemble fini ou dénombrable S avec une variable réelle Y si pour tout $(x, y) \in S \times \mathbb{R}$, nous avons :

$$\mathbb{P}\left(X = x, Y \leq y\right) = \mathbb{P}\left(X = x\right)\mathbb{P}\left(Y \leq y\right).$$

tre 1 :

Définition 13 (Indépendance d'une famille infini de variables aléatoires)**.**

Étant donné $(\Omega, \mathcal{F}, \mathbb{P})$ un espace de probabilité, $\mathcal{I}$ un ensemble de $\mathbb{R}$ fini ou infini (dénombrable comme $\mathbb{N}$ ou non comme $\mathbb{R}$) et $(X_i)_{i\in\mathcal{I}}$ une famille de variables aléatoires, nous disons que les variables sont **indépendantes** si pour toute partie finie I de $\mathcal{I}$ les variables $(X_i)_{i\in\mathcal{I}}$ sont indépendantes.

À l'aide de la notion d'indépendance, nous pouvons introduire la loi du zéro-un de Borel. Avant cela, nous rappelons l'adaptation du théorème de Borel-Cantelli au cas particulier des probabilités :

Lemme 6 (Borel-Cantelli)**.**

Étant donné $(\Omega, \mathcal{F}, \mathbb{P})$ un espace de probabilité et $(A_n)_{n\in\mathbb{N}}$ une suite d'éléments de $\mathcal{F}$, si nous avons

$$\sum_{n\in\mathbb{N}} \mathbb{P}(A_n) < +\infty$$

alors nous avons :

$$\mathbb{P}\left(\limsup_{n\to+\infty} A_n\right) = 0.$$

Preuve

C'est simplement le cas particulier à une mesure de probabilité du théorème de Borel-Cantelli énoncé dans la partie 1.3.1.

Et ainsi, nous avons :

Théorème 7 (Loi du zéro-un de Borel)**.**

Étant donné $(\Omega, \mathcal{F}, \mathbb{P})$ un espace de probabilité et $(A_n)_{n\in\mathbb{N}}$ une suite d'événement alors :

- si la série $\sum_{n\in\mathbb{N}} \mathbb{P}(A_n)$ converge, nous avons $\mathbb{P}(\limsup_{n\to+\infty} A_n) = 0$,
- si les événements sont indépendants et la série $\sum_{n\in\mathbb{N}} \mathbb{P}(A_n)$ divergente, nous avons $\mathbb{P}(\limsup_{n\to+\infty} A_n) = 1$.

Preuve

• Le premier point est directement l'application du lemme de Borel-Cantelli.

• Pour montrer le deuxième point, nous allons calculer la probabilité du complémentaire. Commençons par rappeler ce qu'il représente :

$$\overline{\limsup_{n\to+\infty} A_n} = \overline{\bigcap_{n\in\mathbb{N}} \left(\bigcup_{k\geq n} A_k\right)}$$

$$\begin{aligned} &= \bigcup_{n\in\mathbb{N}} \overline{\left(\bigcup_{k\geq n} A_k\right)} \\ &= \bigcup_{n\in\mathbb{N}} \left(\bigcap_{k\geq n} \overline{A_k}\right). \end{aligned}$$

Pour tout $(n,\ell)\in\mathbb{N}^2$, nous définissons

$$B_n = \bigcap_{k\geq n} \overline{A_k} \text{ et } B_{n,\ell} = \bigcap_{k=n}^{n+\ell} \overline{A_k}$$

et nous avons

$$\forall (n,\ell)\in\mathbb{N}^2, B_{n,\ell+1} = B_{n,\ell}\cap \overline{A_{\ell+1}}$$

donc la suite est décroissante ce qui implique que

$$\forall n\in\mathbb{N}, \lim_{\ell\to+\infty} B_{n,\ell} = B_n.$$

Or, comme les $(A_n)_{n\in\mathbb{N}}$ sont indépendants (donc leurs complémentaires également), nous avons pour tout $(n,\ell)\in\mathbb{N}^2$:

$$\begin{aligned} \mathbb{P}\left(B_{n,\ell}\right) &= \mathbb{P}\left(\bigcap_{k=n}^{n+\ell} \overline{A_k}\right) \\ &= \prod_{k=n}^{n+\ell} \mathbb{P}\left(\overline{A_k}\right) \\ &= \prod_{k=n}^{n+\ell} \left(1-\mathbb{P}\left(A_k\right)\right). \end{aligned}$$

Or, pour tout $t\in[0,1]$, nous avons $0\leq 1-t\leq e^{-t}$ (pour s'en convaincre, il suffit de reprendre la série entière correspondante à $t\mapsto e^t$ ou d'étudier la fonction $t\mapsto 1-t-e^{-t}$) donc nous avons :

$$\begin{aligned} \mathbb{P}\left(B_{n,\ell}\right) &\leq \prod_{k=n}^{n+\ell} e^{-\mathbb{P}(A_k)} \\ &\leq e^{-\sum_{k=n}^{n+\ell}\mathbb{P}(A_k)}. \end{aligned}$$

Or, nous savons par hypothèse que la série des $\sum_{n\in\mathbb{N}}\mathbb{P}\left(A_n\right)$ diverge donc nous avons :

$$\mathbb{P}\left(B_n\right) = \lim_{\ell\to+\infty} \mathbb{P}\left(B_{n,\ell}\right)$$

$$\begin{aligned} &= \lim_{\ell\to+\infty} e^{-\sum_{k=n}^{n+\ell}\mathbb{P}(A_k)} \\ &= \lim_{t\to+\infty} e^{-t} \\ &= 0. \end{aligned}$$

Au final, comme la suite de B_n est croissante, nous avons :

$$\begin{aligned} \mathbb{P}\left(\overline{\limsup_{n\to+\infty} A_n}\right) &= \lim_{n\to+\infty} \mathbb{P}(B_n) \\ &= \lim_{n\to+\infty} 0 \\ &= 0. \end{aligned}$$

Comme la probabilité du complémentaire vaut 0, nous avons le résultat.

⚠ Attention au piège

Notons que l'indépendance est indispensable pour avoir le deuxième résultat. Si nous prenons un événement A tel que $\mathbb{P}(A) = p \in]0;1[$ et la suite d'événements $(A_n)_{n\in\mathbb{N}}$ telle que pour tout $n \in \mathbb{N}$, $A_n = A$, alors la série $\sum_{n\in\mathbb{N}} \mathbb{P}(A_n)$ diverge mais nous avons :

$$\mathbb{P}\left(\limsup_{n\to+\infty} A_n\right) = p.$$

1.4.3 Quantiles des lois de probabilités

Dans cette partie, nous supposons que l'espace de probabilité est $\mathbb{R}$, les définitions peuvent se généraliser à tout espace possédant une relation d'ordre totale.

Définition 14 (Quantile d'ordre p)**.**

Pour tout $p \in]0;1[$, nous appelons **quantile d'ordre** p d'une variable aléatoire X de fonction de répartition F_X est définie par

$$q_p = \inf\{q \in \mathbb{R} \,|\, F_X(q) \geq p\}.$$

En particulier, nous avons :

Définitions 15 (Quantiles particuliers)**.**

- La **médiane** est le quantile d'ordre 50%.
- Les **quartiles** sont les quantiles d'ordre $25k\%$ avec $k \in \{1, 2, 3\}$.
- Les **déciles** sont les quantiles d'ordre $10k\%$ avec $k \in \{1, \dots, 9\}$.
- Les **centiles** sont les quantiles d'ordre $k\%$ avec $k \in \{1, \dots, 99\}$.

Enfin, une caractéristique intéressante est le mode :

Définition 16 (Mode d'une fonction)**.**

Sous condition d'existence, un **mode** d'une variable aléatoire X sur un ensemble S est définie par :

Cas discret	Cas continu
$k^\star \in \underset{k \in S}{\operatorname{argmax}} \mathbb{P}(X = k)$	$x^\star \in \underset{x \in S}{\operatorname{argmax}} f(x)$

⚠ Attention au piège

Il est bien question **d'un** mode car il peut ne pas être unique. En particulier, une loi uniforme sur un ensemble discret possède autant de modes que de valeurs et une loi uniforme sur un intervalle en possède une infinité. De même, l'existence d'un mode n'est pas garantie.
Ce cas peut paraître pathologique mais il existe des fonctions classiques avec plusieurs modes.

Définitions 17 (Loi unimodale, bimodale et multimodale)**.**

Une loi est dite **unimodale** si elle possède un unique mode, **bimodale** si elle en possède deux et **multimodale** si elle en possède plusieurs.

Exercices 1.3

La loi Beta $\mathcal{B}e(a, b)$ de paramètres $(a, b) \in \mathbb{R}_+^\star$ est définie sur $[0; 1]$

par la densité

$$f(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}$$

où $\Gamma(\cdot)$ est la fonction définie sur $\mathbb{R}_+^\star$ par :

$$\Gamma(x) = \int_0^{+\infty} t^{x-1} e^{-x}.$$

Déterminer le nombre de modes de la loi en fonction de a et b.

1.5 Vecteurs gaussiens

Dans cette partie, nous commençons par rappeler quelques propriétés importantes des lois gaussiennes en une dimension puis nous généralisons aux vecteurs gaussiens.

1.5.1 Variable gaussienne unidimensionnelle

Pour toute cette partie, nous nous placerons sur $\mathbb{R}$ muni de la tribu borélienne.

Définition 18 (Variable gaussienne centrée réduite)**.**

Une variable aléatoire X suit une **loi gaussienne** (ou **normale**) **centrée réduite** si sa densité de probabilité est :

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}.$$

Une représentation de la densité est disponible dans la figure 1.1. Nous notons cette loi $\mathcal{N}(0,1)$.

La loi gaussienne centrée réduite est l'une des plus étudiées car elle revient dans de nombreux cas.

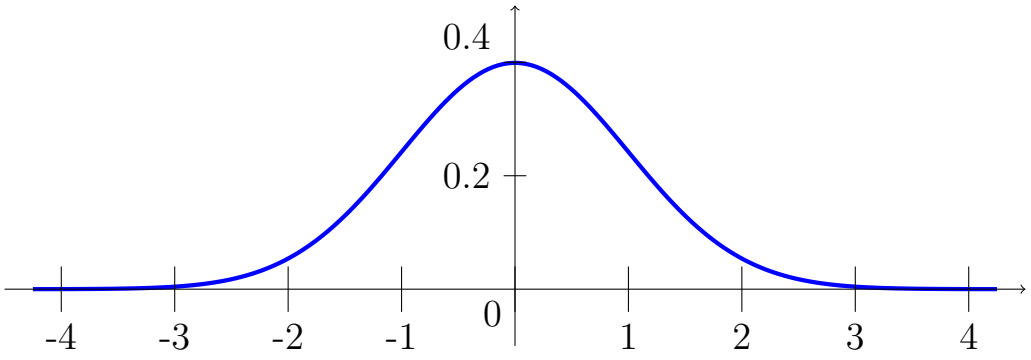


FIGURE 1.1 – Représentation de la densité d'une loi gaussienne centrée réduite.

Propriétés 8 (Symétrie de la loi gaussienne)**.**

La loi gaussienne centrée est **symétrique par rapport à l'origine** c'est-à-dire que sa densité est paire :

$$\forall x \in \mathbb{R},\ f(-x) = f(x).$$

En particulier, si X est une variable gaussienne centrée alors, pour tout $x \in \mathbb{R}$, nous avons donc :

$$\mathbb{P}(X \geq x) = \mathbb{P}(X \leq -x).$$

Preuve

• La symétrie par rapport à l'origine vient de la parité de la fonction carré $x \mapsto x^2$.

• Pour la deuxième partie, il suffit de faire le calcul pour tout $x \in \mathbb{R}$:

$$\begin{aligned}
\mathbb{P}(X \geq x) &= \int_x^{+\infty} f(t)\mathrm{d}t \\
u = -t, \qquad t = -u \text{ et } \frac{\mathrm{d}t}{\mathrm{d}u} &= -1, \\
&= \int_{-x}^{-\infty} f(-u) \times (-1)\mathrm{d}u \\
&= \int_{-\infty}^{-x} f(u)\mathrm{d}u \\
&= \mathbb{P}(X \leq -x).
\end{aligned}$$

Remarque

La deuxième probabilité implique que n'avons besoin de connaître

que les quantiles de la loi pour des valeurs positives. Une représentation schématique de la propriété est proposée sur la figure 1.2.

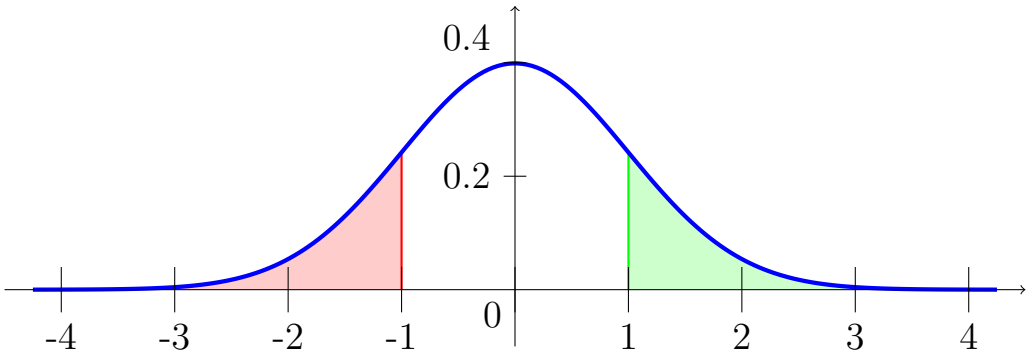


FIGURE 1.2 – Représentation de la densité d'une loi gaussienne centrée réduite : la zone rouge à gauche a la même aire que la zone verte à droite.

⚠ Attention au piège

La fonction de répartition d'une loi gaussienne ne peut pas être calculée de façon analytique pour presque tous les t (excepté pour $t = 0$ où elle vaut $1/2$). Le tableau 1.1 représente la table de loi de la loi gaussienne centrée réduite.

Proposition 9 (Moments d'ordre k d'une loi gaussienne centrée réduite)**.**

Les moments d'ordre k d'une variable aléatoire X suivant une loi gaussienne centrée réduite sont :

$$\forall k \in \mathbb{N},\ \mathbb{E}\left[X^k\right] = \begin{cases} \frac{(2m)!}{2^m m!} & \text{si } k = 2m, \\ 0 & \text{sinon.} \end{cases}$$

Preuve

Commençons par remarquer que tous les moments sont bien définis en faisant une comparaison à n'importe quelle fonction polynomiale de référence car la fonction exponentielle décroît bien plus vite que les fonctions polynomiales. Pour les moments d'ordre impair, cela vient de la symétrie de la loi. Pour les autres, nous avons pour tout $m \in \mathbb{N}^\star$, la relation suivante :

$$\mathbb{E}\left[X^{2m}\right] = \int_{\mathbb{R}} x^{2m} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \mathrm{d}x$$

$$
\begin{array}{rlrl}
u(x) &= x^{2m-1} & u'(x) &= (2m-1)x^{2m-2} \\
v'(x) &= \dfrac{x}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} & v(x) &= -\dfrac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}
\end{array}
$$

$$
\begin{aligned}
&= \left[-\frac{x^{2m-1}}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \right]_{-\infty}^{+\infty} - \int_{\mathbb{R}} \frac{-(2m-1)x^{2m-2}}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \mathrm{d}x \\
&= 0 - 0 + \int_{\mathbb{R}} \frac{(2m-1)x^{2m-2}}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \mathrm{d}x \\
&= (2m-1)\mathbb{E}\left[X^{2m-2}\right].
\end{aligned}
$$

Nous procédons donc par récurrence :

$$
\begin{aligned}
\mathbb{E}\left[X^0\right] &= 1, \\
\mathbb{E}\left[X^2\right] &= 1 \times 1 \\
&= 1, \\
\mathbb{E}\left[X^4\right] &= 3, \\
&\vdots \\
\mathbb{E}\left[X^{2m}\right] &= 1 \times 3 \times \cdots \times (2m-1) \\
&= \frac{1 \times 2 \times 3 \times 4 \times \cdots \times (2m-1) \times (2m)}{2 \times 4 \times \cdots \times (2m)} \\
&= \frac{(2m)!}{2^m m!}.
\end{aligned}
$$

Définitions 19 (Loi gaussienne)**.**

Pour tout couple $(\mu, \sigma) \in \mathbb{R} \times \mathbb{R}^+$, une variable aléatoire X suit une **loi gaussienne** (ou **normale**) **de moyenne μ et de variance σ^2** s'il existe une variable Z gaussienne centrée réduite telle que :

$$X = \mu + \sigma Z.$$

Nous notons cette loi $\mathcal{N}(\mu, \sigma^2)$.
Si $\sigma = 0$, nous dirons que la variable gaussienne est **dégénérée**.

⚠ Attention au piège

Certains auteurs notent la loi avec l'écart-type plutôt que la va-

riance (donc avec $\mathcal{N}(\mu, \sigma)$ au lieu de $\mathcal{N}(\mu, \sigma^2)$). Dans ce cours, nous avons fait le choix retenu par la majorité des lectures proposées dans la bibliographie.

Propriétés 10 (Densité d'une loi gaussienne non dégénérée)**.**

Pour tout couple $(\mu, \sigma) \in \mathbb{R} \times \mathbb{R}_+^\star$, la loi gaussienne $\mathcal{N}(\mu, \sigma^2)$ a pour densité :

$$f_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}.$$

Preuve

La preuve est simplement un changement de variable possible uniquement car $\sigma > 0$. Pour tout couple $(\mu, \sigma) \in \mathbb{R} \times \mathbb{R}_+^\star$, si X suit une loi gaussienne $\mathcal{N}(\mu, \sigma^2)$ alors il existe Z loi gaussienne centrée réduite telle que pour tout fonction mesurable h, nous ayons :

$$\begin{aligned}
\mathbb{E}\left[h(X)\right] &= \mathbb{E}\left[h(\mu + \sigma Z)\right] \\
&= \int_{-\infty}^{+\infty} h(\mu + \sigma z) \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \mathrm{d}z \\
x = \mu + \sigma z, \qquad z = \frac{x-\mu}{\sigma} \text{ et } \frac{\mathrm{d}z}{\mathrm{d}x} &= \frac{1}{\sigma} \\
&= \int_{-\infty}^{+\infty} h(x) \frac{1}{\sqrt{2\pi}} e^{-\frac{\left(\frac{x-\mu}{\sigma}\right)^2}{2}} \frac{\mathrm{d}x}{\sigma} \\
&= \int_{-\infty}^{+\infty} h(x) \underbrace{\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}}_{=f_{\mu,\sigma^2}(x)} \mathrm{d}x.
\end{aligned}$$

Définition 20 (Fonction caractéristique)**.**

Étant donnée une variable aléatoire X à valeur dans $\mathbb{R}^d$, la **fonction caractéristique de** X, notée Φ_X, est définie pour tout $\xi \in \mathbb{R}^d$ par :

$$\begin{aligned}
\Phi_X : \mathbb{R}^d &\to \mathbb{C} \\
\xi &\mapsto \mathbb{E}\left[e^{\langle i\xi, X\rangle}\right]
\end{aligned}$$

où i est le nombre imaginaire tel que $i^2 = -1$ et $\langle \cdot, \cdot \rangle$ est le produit scalaire.

Remarque

La fonction caractéristique peut être vue comme la transformée de Fourier de la loi de X ce qui permet d'utiliser tous les résultats classiques dans ce domaine (continuité de la fonction caractéristique par exemple).

Proposition 11 (Fonction caractéristique d'une loi gaussienne)**.**

Pour tout couple $(\mu, \sigma) \in \mathbb{R} \times \mathbb{R}^+$, la fonction caractéristique de la loi gaussienne $\mathcal{N}(\mu, \sigma^2)$ est définie pour tout $\xi \in \mathbb{R}$ par :

$$\Phi_{\mu,\sigma^2}(\xi) = e^{i\xi\mu - \frac{\sigma^2\xi^2}{2}}.$$

Preuve

Commençons par montrer le résultat pour la loi gaussienne centrée réduite.

Lemme 12 (Fonction caractéristique d'une loi gaussienne centrée réduite)**.**

La fonction caractéristique de la loi gaussienne centrée réduite est définie pour tout $\xi \in \mathbb{R}$ par :

$$\Phi_{0,1}(\xi) = e^{-\frac{\xi^2}{2}}.$$

Preuve

Pour tout $\xi \in \mathbb{R}$, nous avons :

$$\begin{aligned}
\Phi_{0,1}(\xi) &= \mathbb{E}\left[e^{i\xi Z}\right] \text{ où } Z \sim \mathcal{N}(0,1) \\
&= \int_{\mathbb{R}} e^{i\xi z} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \mathrm{d}z \\
&= \int_{\mathbb{R}} \cos(\xi z) \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \mathrm{d}z + i \int_{\mathbb{R}} \underbrace{\sin(\xi z) \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}}_{\text{fonction impaire en } z} \mathrm{d}z \\
&= \int_{\mathbb{R}} \cos(\xi z) \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \mathrm{d}z.
\end{aligned}$$

La dérivée de $\xi \mapsto \cos(\xi z)\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}$ est $\xi \mapsto -z\sin(\xi z)\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}$ pour tout $\xi \in \mathbb{R}$ et $z \in \mathbb{R}$

$$\left| -z \sin(\xi z) \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \right| \leq \max(|z|, 1) e^{-z^2}$$

avec $z \mapsto \max(|z|,1)e^{-z^2}$ intégrable. Donc, par le théorème de dérivation sous l'intégrale, nous avons donc :

$$\begin{aligned}
\Phi'_{0,1}(\xi) &= \int_{\mathbb{R}} -z \sin(\xi z) \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \mathrm{d}z \\
&= -\int_{\mathbb{R}} \sin(\xi z) \frac{z}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \mathrm{d}z \\
u(z) = \sin(\xi z) \qquad & u'(z) = \xi \cos(\xi z) \\
v'(z) = \frac{z}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \qquad & v(z) = -\frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}
\end{aligned}$$

$$\begin{aligned} &= -\left(\left[-\sin(\xi z)\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}\right]_{\infty}^{+\infty} + \int_{-\infty}^{+\infty}\xi\cos(\xi z)\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}\mathrm{d}z\right) \\ &= -\xi\int_{-\infty}^{+\infty}\cos(\xi z)\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}\mathrm{d}z \\ &= -\xi\Phi_{0,1}(\xi). \end{aligned}$$

Par conséquent, $\Phi_{0,1}$ vérifie l'équation différentielle $\Phi'_{0,1}(\xi) = -\xi\Phi_{0,1}(\xi)$. Or, une primitive de $\xi \mapsto \xi$ est $\xi \mapsto \frac{\xi^2}{2}$ donc $\Phi_{0,1}$ est de la forme :

$$\Phi_{0,1}(\xi) = ce^{-\frac{\xi^2}{2}} \text{ avec } c \in \mathbb{R}.$$

Comme nous avons

$$c = ce^{-\frac{0^2}{2}} = \Phi_{0,1}(0) = \mathbb{E}\left[e^{i0Z}\right] = \mathbb{E}\left[1\right] = 1,$$

alors, pour la loi gaussienne centrée réduite, nous avons $\Phi_{0,1}(\xi) = e^{-\frac{\xi^2}{2}}$.

Pour la généralisation, nous savons que pour tout couple $(\mu, \sigma) \in \mathbb{R} \times \mathbb{R}^+$, si X suit une loi $\mathcal{N}(\mu, \sigma^2)$ alors il existe Z suivant une loi gaussienne centrée réduite telle que $X = \mu + \sigma Z$ donc nous avons :

$$\begin{aligned} \Phi_{\mu,\sigma^2}(\xi) &= \mathbb{E}\left[e^{i\xi X}\right] \\ &= \mathbb{E}\left[e^{i\xi(\mu+\sigma Z)}\right] \\ &= \mathbb{E}\left[e^{i\xi\mu}e^{i\xi\sigma Z}\right] \\ &= e^{i\xi\mu}\mathbb{E}\left[e^{i(\xi\sigma)Z}\right] \\ &= e^{i\xi\mu}\Phi_{0,1}(\xi\sigma) \\ &= e^{i\xi\mu}e^{-\frac{(\xi\sigma)^2}{2}} \\ &= e^{i\xi\mu-\frac{\xi^2\sigma^2}{2}}. \end{aligned}$$

1.5.2 Vecteur gaussien

Dans cette partie, nous nous plaçons dans $\mathbb{R}^d$ munie de la tribu borélienne.

Définitions 21 (Vecteur gaussien)**.**

Un vecteur aléatoire $\mathbf{X}$ à valeurs dans $\mathbb{R}^d$ est dit **gaussien** si toute combinaison linéaire de ses composantes suit une loi gaussienne.
Si $\mathbf{X} = (X_1, \ldots, X_d)^T$ (où A^T symbolise la **transposée** de la matrice A) est un vecteur gaussien, on définit son **vecteur moyenne** (ou simplement sa **moyenne**) $\mathbb{E}[\mathbf{X}]$ par $\mathbb{E}[\mathbf{X}] = (\mathbb{E}[X_1], \ldots, \mathbb{E}[X_d])^T$ et matrice de **variance-covariance** par $\mathbb{V}[\mathbf{X}] = \mathbb{E}\left[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^T\right]$.

Propriétés 13 (Matrice de variance-covariance)**.**

La matrice de variance covariance est symétrique, positive et pour tout $(i,j) \in \{1, \ldots, d\}^2$, nous avons $(\mathbb{V}[\mathbf{X}])_{i,j} = \mathrm{Cov}(X_i, X_j)$.

Preuve

- La symétrie vient de la commutativité de la multiplication.
- Pour la positivité, prenons $u \in \mathbb{R}^d$ et calculons :

$$\begin{aligned} u^T \mathbb{V}[\mathbf{X}]\, u &= \sum_{i=1}^{d}\sum_{j=1}^{d} u_i \mathrm{Cov}(X_i, X_j)\, u_j \\ &= \sum_{i=1}^{d}\sum_{j=1}^{d} \mathrm{Cov}(u_i X_i, u_j X_j) \\ &= \sum_{i=1}^{d} \mathrm{Cov}\left(u_i X_i, \sum_{j=1}^{d} u_j X_j\right) \\ &= \mathrm{Cov}\left(\sum_{i=1}^{d} u_i X_i, \sum_{j=1}^{d} u_j X_j\right) \\ &= \mathbb{V}\left[\sum_{i=1}^{d} u_i X_i\right] \\ &\geq 0. \end{aligned}$$

Comme ceci est vrai pour tout $u \in \mathbb{R}^d$, nous avons la positivité.
- Pour le dernier point, c'est la définition du produit de deux vecteurs.

Remarque

Comme n'importe quelle combinaison linéaire des coordonnées de $\mathbf{X}$ sont gaussiennes alors chaque coordonnée, en particulier, l'est : pour

cela, il suffit d'associer le vecteur de $\mathbb{R}^d$ valant 0 partout sauf pour une coordonnée.

⚠ Attention au piège

La réciproque est fausse.

Contre-exemple

Si nous prenons Z une variable gaussienne centrée réduite et X une variable indépendante de Z suivant une loi de Bernoulli sur $\{-1,1\}$ de paramètre $p \in]0,1[$ alors le vecteur (Z, XZ) a ses deux coordonnées qui sont gaussiennes mais ce n'est pas un vecteur gaussien puisque $Z+XZ$ vaudra 0 avec une probabilité $(1-p)$ et suivra une loi gaussienne $\mathcal{N}(0,4)$ avec probabilité p.

Théorème 14 (Fonction caractéristique)**.**

Soit $\mathbf{X} = (X_1, \ldots, X_d)^T$ un vecteur gaussien d'espérance $m = \mathbb{E}[\mathbf{X}]$ et de matrice de variance covariance $\Sigma = \mathbb{V}[\mathbf{X}]$ alors nous avons pour tout $\xi \in \mathbb{R}^d$:

$$\Phi_{\mathbf{X}}(\xi) = \mathbb{E}\left[e^{i\langle \xi, \mathbf{X}\rangle}\right] = e^{i\xi^T m - \frac{\xi^T \Sigma \xi}{2}}.$$

La loi de $\mathbf{X}$ étant entièrement déterminée par m et Σ, nous l'appelons la **loi gaussienne** (ou **normale**) **de paramètre** m **et** Σ notée $\mathcal{N}(m, \Sigma)$.

Preuve

Par définition d'un vecteur gaussien, pour tout $\xi \in \mathbb{R}^d$, la variable $\langle \xi, \mathbf{X}\rangle$ est une combinaison linéaire des coordonnées de $\mathbf{X}$ donc est gaussienne de moyenne $\mathbb{E}\left[\xi^T \mathbf{X}\right] = \xi^T m$ et de variance $\mathbb{V}\left[\xi^T \mathbf{X}\right] = \xi^T \Sigma \xi$. Donc, par la proposition 11, nous avons pour tout $u \in \mathbb{R}$:

$$\Phi_{\xi^T \mathbf{X}}(u) = e^{iu\xi^T m - \frac{\xi^T \Sigma \xi}{2} u^2}.$$

Au final, nous avons donc pour tout $\xi \in \mathbb{R}^d$:

$$\begin{aligned} \Phi_{\mathbf{X}}(\xi) &= \mathbb{E}\left[e^{i\xi^T \mathbf{X}}\right] \\ &= \Phi_{\xi^T \mathbf{X}}(1) \\ &= e^{i\xi^T m - \frac{\xi^T \Sigma \xi}{2}}. \end{aligned}$$

À l'aide de ce théorème, nous pouvons prouver les deux propositions suivantes :

Proposition 15 (Linéarité des vecteurs gaussiens)**.**

Étant donnée $\mathbf{X} = (X_1, \ldots, X_d)$ un vecteur gaussien d'espérance m et de matrice de variance covariance Σ alors pour toute matrice $A \in \mathcal{M}_{k\times d}(\mathbb{R})$ et pour tout vecteur $b \in \mathbb{R}^k$, le vecteur $A\mathbf{X} + b$ est gaussien de loi $\mathcal{N}(Am + b, A\Sigma A^T)$.

Preuve

Pour cela, il suffit de calculer la fonction caractéristique, c'est-à-dire pour tout $\xi \in \mathbb{R}^k$:

$$\begin{aligned}
\Phi_{A\mathbf{X}+b}(\xi) &= \mathbb{E}\left[e^{i\langle \xi, A\mathbf{X}+b\rangle}\right] \\
&= \mathbb{E}\left[e^{i\langle \xi, A\mathbf{X}\rangle} e^{i\langle \xi, b\rangle}\right] \\
&= \mathbb{E}\left[e^{i\xi^T A\mathbf{X}} e^{i\xi^T b}\right] \\
&= \mathbb{E}\left[e^{i\left(\left(A^T\xi\right)^T\right)\mathbf{X}} e^{i\xi^T b}\right] \\
&= e^{i\xi^T b}\Phi_{\mathbf{X}}\left(A^T\xi\right) \\
&= e^{i\xi^T b} e^{i\left(A^T\xi\right)^T m - \frac{\left(A^T\xi\right)^T \Sigma \left(A^T\xi\right)}{2}} \\
&= e^{i\xi^T b + i\xi^T Am - \frac{\xi^T A\Sigma A^T \xi}{2}} \\
&= e^{i\xi^T(Am+b) - \frac{\xi^T\left(A\Sigma A^T\right)\xi}{2}}.
\end{aligned}$$

Ce qui permet de conclure.

Proposition 16 (Indépendance des coordonnées d'un vecteur gaussien)**.**

Étant donnée $\mathbf{X} = (X_1, \ldots, X_d)$ un vecteur gaussien alors pour tout $(k, \ell) \in \{1, \ldots, d\}$ avec k différent de ℓ, nous avons :

$$X_k \text{ et } X_\ell \text{ sont indépendants} \Leftrightarrow \mathrm{Cov}\left(X_k, X_\ell\right) = 0.$$

Preuve

Sens $\Rightarrow$: ce sens est simplement une propriété de la covariance.

Sens $\Leftarrow$: Prenons k et ℓ différents dans $\{1,\dots,d\}$ et supposons que $\mathrm{Cov}(X_k, X_\ell) = 0$ et la matrice P de projection du vecteur $\mathbf{X}$ sur le vecteur $(X_k, X_\ell)^T$; c'est à dire que nous avons :

$$P = \begin{pmatrix} 0 & \cdots & 0 & 1 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \end{pmatrix}$$

où les 1 sont respectivement en $k^{\text{ème}}$ et $\ell^{\text{ème}}$ position (dans cet exemple, nous avons supposé que $k < \ell$). Par la proposition 15, nous savons que $(X_k, X_\ell)^T = P\mathbf{X}$ est gaussien de moyenne $\mathbb{E}\left[(X_k, X_\ell)^T\right] = (\mathbb{E}[X_k], \mathbb{E}[X_\ell])^T$ et de matrice de variance covariance :

$$\begin{aligned} \mathbb{V}\left[\begin{pmatrix} X_k \\ X_\ell \end{pmatrix}\right] &= P\left(\mathrm{Cov}(X_{k'}, X_{\ell'})\right)_{k',\ell'} P^T \\ &= \begin{pmatrix} \mathrm{Cov}(X_k, X_1) & \cdots \mathrm{Cov}(X_k, X_d) \\ \mathrm{Cov}(X_\ell, X_1) & \cdots \mathrm{Cov}(X_\ell, X_d) \end{pmatrix} P^T \\ &= \begin{pmatrix} \mathrm{Cov}(X_k, X_k) & \mathrm{Cov}(X_k, X_\ell) \\ \mathrm{Cov}(X_\ell, X_k) & \mathrm{Cov}(X_\ell, X_\ell) \end{pmatrix} \\ &= \begin{pmatrix} \mathbb{V}[X_k] & 0 \\ 0 & \mathbb{V}[X_\ell] \end{pmatrix}. \end{aligned}$$

Donc, pour tout $(\xi_1, \xi_2) \in \mathbb{R}^2$, nous avons :

$$\begin{aligned} \Phi_{\begin{pmatrix} X_k \\ X_\ell \end{pmatrix}}\left(\begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix}\right) &= \exp\left\{ i \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix}^T \begin{pmatrix} X_k \\ X_\ell \end{pmatrix} - \frac{1}{2} \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix}^T \begin{pmatrix} \mathbb{V}[X_k] & 0 \\ 0 & \mathbb{V}[X_\ell] \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} \right\} \\ &= e^{i(\xi_1 X_k + \xi_2 X_\ell) + \frac{\mathbb{V}[X_k]\xi_1^2 + \mathbb{V}[X_\ell]\xi_2^2}{2}} \\ &= e^{i\xi_1 X_k + \frac{\mathbb{V}[X_k]\xi_1^2}{2}} e^{i\xi_2 X_\ell + \frac{\mathbb{V}[X_\ell]\xi_2^2}{2}} \\ &= \Phi_{X_k}(\xi_1)\Phi_{X_\ell}(\xi_2) \end{aligned}$$

ce qui prouve l'indépendance.

⚠ Attention au piège

L'équivalence n'est possible que parce que nous avons un vecteur

gaussien.

Contre-exemple

Si nous prenons Z une variable gaussienne centrée réduite et X une variable indépendante de Z suivant une loi de Bernoulli sur $\{-1,1\}$ de paramètre $p=1/2$ alors Z et XZ sont dépendants mais $\mathrm{Cov}(Z,XZ)=0$.

Théorème 17 (Densité d'un vecteur gaussien)**.**

Étant donné $\mathbf{X}=(X_1,\dots,X_d)^T$ un vecteur gaussien d'espérance $m=\mathbb{E}[\mathbf{X}]$ et de matrice de variance covariance $\Sigma=\mathbb{V}[\mathbf{X}]$, la loi de $\mathbf{X}$ admet une densité $f_{m,\Sigma}$ par rapport à la mesure de Lebesgue sur $\mathbb{R}^d$ si et seulement si Σ est inversible. Dans ce cas, nous avons pour tout $\mathbf{x}\in\mathbb{R}^d$,

$$f_{m,\Sigma}(\mathbf{x})=\frac{1}{\sqrt{2\pi}^d}\frac{1}{\sqrt{|\Sigma|}}e^{-\frac{1}{2}(\mathbf{x}-m)^T\Sigma^{-1}(\mathbf{x}-m)}$$

où $|\Sigma|$ représente le déterminant de la matrice Σ.
Si, en revanche, Σ n'est pas inversible, la loi de $\mathbf{X}-m$ est presque sûrement portée par le sous-espace vectoriel engendré par les vecteurs propres aux valeurs propres non nulles de Σ.

Preuve

D'après les propriétés 13 de la matrice Σ, nous savons qu'elle est symétrique et positive donc elle est diagonalisable dans une base orthonormée de vecteurs propres $u_1,\dots,u_d$ associés aux valeurs propres $\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_d\geq 0$. On note U la matrice orthogonale formée par les vecteurs u_j et r le rang de Σ ; c'est-à-dire que $\lambda_r>$ et $\lambda_{r+1}=\lambda_{r+2}=\cdots=\lambda_d=0$.

On peut réécrire la matrice de variance covariance comme $\Sigma=U\Gamma^{-1}U^T=\left(U\sqrt{\Gamma}\right)\left(U\sqrt{\Gamma}\right)^T$ avec

$$\Gamma=\begin{pmatrix}\lambda_1 & 0 & \cdots & 0 & \\ 0 & \lambda_2 & \ddots & \vdots & \\ \vdots & \ddots & \ddots & 0 & \mathbf{0}_{r\times(d-r)} \\ 0 & \cdots & 0 & \lambda_r & \\ & \mathbf{0}_{(d-r)\times r} & & & \mathbf{0}_{(d-r)\times(d-r)}\end{pmatrix}\text{ et }\sqrt{\Gamma}=\begin{pmatrix}\sqrt{\lambda_1} & 0 & \cdots & 0 & \\ 0 & \sqrt{\lambda_2} & \ddots & \vdots & \\ \vdots & \ddots & \ddots & 0 & \mathbf{0}_{r\times(d-r)} \\ 0 & \cdots & 0 & \sqrt{\lambda_r} & \\ & \mathbf{0}_{(d-r)\times r} & & & \mathbf{0}_{(d-r)\times(d-r)}\end{pmatrix}$$

Par la proposition 15, il existe un vecteur gaussien $\mathbf{Y}$ de loi $\mathcal{N}(0,\mathbb{I}_d)$

tel que $\mathbf{X}$ a même loi que $m + U\sqrt{\Gamma}\mathbf{Y}$. Comme la matrice $U\sqrt{\Gamma}$ permet d'avoir une combinaison linéaire des r premières coordonnées du vecteur $\mathbf{X}$ alors si Σ n'est pas inversible ($r < n$), la loi de $\mathbf{X}_m$ admet pour support le sous espace vectoriel de $\mathbb{R}^d$ engendré par $u_1, \ldots, u_r$.
Sinon, nous utilisons le lemme suivant :

Lemme 18 (Fonction caractéristique d'une loi gaussienne $\mathcal{N}(0, \mathbb{I}_d)$)**.**

La densité d'un vecteur gaussien de loi $\mathcal{N}(0, \mathbb{I}_d)$ est définie pour tout $\mathbf{y} \in \mathbb{R}^d$ par :

$$f_{0,\mathbb{I}_d}(\mathbf{y}) = \frac{1}{\sqrt{2\pi}^d} e^{-\frac{1}{2}\mathbf{y}^T\mathbf{y}}.$$

Preuve

Si $\mathbf{Y}$ est un vecteur gaussien suivant une loi $\mathcal{N}(0, \mathbb{I}_d)$ alors par définition d'un vecteur gaussien et par la propriété 16 d'indépendance, les coordonnées du vecteur sont indépendantes et de même loi $\mathcal{N}(0, 1)$ donc la densité de la loi jointe des coordonnées (donc du vecteur) est le produit des densité ; c'est-à-dire que pour tout $\mathbf{y} \in \mathbb{R}^d$, nous avons :

$$\begin{aligned} f_{0,\mathbb{I}_d}(\mathbf{y}) &= \prod_{i=1}^{d} f_{0,1}(y_i) \\ &= \prod_{i=1}^{d} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y_i^2} \\ &= \frac{1}{\sqrt{2\pi}^d} e^{-\frac{1}{2}\sum_{i=1}^{d} y_i^2} \\ &= \frac{1}{\sqrt{2\pi}^d} e^{-\frac{1}{2}\mathbf{y}^T\mathbf{y}}. \end{aligned}$$

À partir de ce lemme, nous avons donc pour toute fonction mesurable h de $\mathbb{R}^d$ vers $\mathbb{R}$:

$$\begin{aligned}
\mathbb{E}\left[h(\mathbf{X})\right] &= \mathbb{E}\left[h\left(m+U\sqrt{\Gamma}\mathbf{Y}\right)\right] \\
&= \iint_{\mathbb{R}^d} h\left(m+U\sqrt{\Gamma}\mathbf{y}\right)\frac{1}{\sqrt{2\pi}^d}e^{-\frac{1}{2}\mathbf{y}^T\mathbf{y}}\mathrm{d}\mathbf{y} \\
\mathbf{x}=m+U\sqrt{\Gamma}\mathbf{y}, \qquad & \mathbf{y}=\left(U\sqrt{\Gamma}\right)^{-1}(\mathbf{x}-m) \text{ et } \mathrm{d}\mathbf{y}=\left|\left(U\sqrt{\Gamma}\right)^{-1}\right|\mathrm{d}\mathbf{x} \\
&= \iint_{\mathbb{R}^d} h(\mathbf{x})\frac{1}{\sqrt{2\pi}^d}e^{-\frac{1}{2}\left[\left(U\sqrt{\Gamma}\right)^{-1}(\mathbf{x}-m)\right]^T\left[\left(U\sqrt{\Gamma}\right)^{-1}(\mathbf{x}-m)\right]}\left|\left(U\sqrt{\Gamma}\right)^{-1}\right|\mathrm{d}\mathbf{x} \\
&= \iint_{\mathbb{R}^d} h(\mathbf{x})\frac{1}{\sqrt{2\pi}^d}e^{-\frac{1}{2}(\mathbf{x}-m)^T\left[\left(U\sqrt{\Gamma}\right)^{-1}\right]^T\left(U\sqrt{\Gamma}\right)^{-1}(\mathbf{x}-m)}\left|\left(\sqrt{\Gamma}\right)^{-1}U^{-1}\right|\mathrm{d}\mathbf{x} \\
&= \iint_{\mathbb{R}^d} h(\mathbf{x})\frac{1}{\sqrt{2\pi}^d}e^{-\frac{1}{2}(\mathbf{x}-m)^T\left[\left(U\sqrt{\Gamma}\right)^T\right]^{-1}\left(U\sqrt{\Gamma}\right)^{-1}(\mathbf{x}-m)}\left|\left(\sqrt{\Gamma}\right)^{-1}\right|\underbrace{\left|U^{-1}\right|}_{=1}\mathrm{d}\mathbf{x} \\
&= \iint_{\mathbb{R}^d} h(\mathbf{x})\frac{1}{\sqrt{2\pi}^d}e^{-\frac{1}{2}(\mathbf{x}-m)^T\left[\sqrt{\Gamma}U^T\right]^{-1}\left(U\sqrt{\Gamma}\right)^{-1}(\mathbf{x}-m)}\prod_{i=1}^{n}\frac{1}{\sqrt{\lambda_i}}\mathrm{d}\mathbf{x} \\
&= \iint_{\mathbb{R}^d} h(\mathbf{x})\frac{1}{\sqrt{2\pi}^d}e^{-\frac{1}{2}(\mathbf{x}-m)^T\left(U\sqrt{\Gamma}\sqrt{\Gamma}U^T\right)^{-1}(\mathbf{x}-m)}\frac{1}{\prod_{i=1}^{n}\sqrt{\lambda_i}}\mathrm{d}\mathbf{x} \\
&= \iint_{\mathbb{R}^d} h(\mathbf{x})\frac{1}{\sqrt{2\pi}^d}e^{-\frac{1}{2}(\mathbf{x}-m)^T\left(U\Gamma U^T\right)^{-1}(\mathbf{x}-m)}\frac{1}{\sqrt{\prod_{i=1}^{n}\lambda_i}}\mathrm{d}\mathbf{x} \\
&= \iint_{\mathbb{R}^d} h(\mathbf{x})\frac{1}{\sqrt{2\pi}^d}e^{-\frac{1}{2}(\mathbf{x}-m)^T\Sigma^{-1}(\mathbf{x}-m)}\frac{1}{\sqrt{|\Sigma|}}\mathrm{d}\mathbf{x}.
\end{aligned}$$

Ensuite, nous voyons une proposition importante pour la statistique (mais malheureusement hors programme puisque les espérances conditionnelles n'ont pas encore été vues).

Proposition 19 (Espérance conditionnelle)**.**

Si $(Y, X_1, \ldots, X_d)^T$ est un vecteur gaussien alors $\mathbb{E}\left[Y\,|X_1, \ldots, X_d\right]$ est une fonction affine des composantes de $\mathbf{X} = (X_1, \ldots, X_d)^T$.

Preuve

Si nous notons $F = \mathrm{Vect}\,(1, X_1, \ldots, X_d)$ l'espace vectoriel engendré

par le vecteur gaussien $(1, \mathbf{X})$ et $p_F(Y)$ la projection de Y sur pour le produit scalaire associé à l'espérance (c'est-à-dire $\langle X, Y\rangle_{\mathbb{E}[\cdot]} = \mathbb{E}[XY]$) alors pour toute variable $Z \in F$, nous avons par définition :

$$\mathbb{E}[(Y - p_F(Y))Z] = 0.$$

Si nous prenons $Z = 1$ alors nous en déduisons que $\mathbb{E}[Y - p_F(Y)] = 0$. De même, si nous prenons $Z = X_j$ avec $j \in \{1, \ldots, d\}$ alors

$$\begin{aligned} 0 &= \mathbb{E}[(Y - p_F(Y))X_j] \\ &= \mathrm{Cov}(Y - p_F(Y), X_j). \end{aligned}$$

Donc, $Y - p_F(Y)$ et X_j sont indépendantes. Comme le vecteur $(Y - p_F(Y), X_1, \ldots, X_d)$ est gaussien alors $Y - p_F(Y)$ est indépendante de toute fonction de $(X_1, \ldots, X_d)$ et $p_F(y) = \mathbb{E}[Y \,|X_1, \ldots, X_d]$.

Enfin, nous pouvons montrer le théorème de la limite centrale dans le cas multidimensionnelle.

Théorème 20 (Version multidimensionnelle de la limite centrale)**.**

Étant donné $(Y_1, \ldots, Y_n)$ une suite de vecteurs de $\mathbb{R}^d$ indépendants et de même loi admettant un moment d'ordre 2 dont on note l'espérance m et la matrice de variance covariance Σ communes alors

$$\sqrt{n}\left(\overline{Y}_n - m\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma).$$

Lemme 21 (Version unidimensionnelle du théorème de limite centrale)**.**

Étant donné $(Y_1, \ldots, Y_n)$ une suite de vecteurs de $\mathbb{R}^d$ indépendants et de même loi admettant un moment d'ordre 2 dont on note l'espérance m et σ^2 la variance communes alors

$$\sqrt{n}\left(\overline{Y}_n - m\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2)$$

avec $\overline{Y}_n = \frac{1}{n}\sum_{i=1}^n Y_i$.

Preuve

Commençons par noter $Z_n = \sqrt{n}\left(\overline{Y_n} - m\right)$ et montrons la conver-

gence simple de la fonction caractéristique. Pour tout $\xi \in \mathbb{R}^n$, $\left(\xi^T V_1, \ldots, \xi^T V_n\right)$ est un échantillon indépendant et identiquement distribué dans $\mathbb{R}$ de moyenne $\xi^T m$ et de variance $\xi^T \Sigma \xi$ donc par le théorème de limite centrale unidimensionnel, nous avons :

$$\xi^T Z_n = \sqrt{n}\left(\overline{\xi^T Z_n} - \xi^T m\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \xi^T \Sigma \xi\right).$$

Une autre façon de le dire est que pour tout $t \in \mathbb{R}$:

$$\Phi_{\xi^T Z_n}(t) \underset{n\to+\infty}{\longrightarrow} \exp\left(-\frac{1}{2}\xi^T \Sigma \xi t^2\right).$$

Donc, au final, nous avons :

$$\Phi_{Z_n}(\xi) = \Phi_{\xi^T Z_n}(1) \underset{n\to+\infty}{\longrightarrow} \exp\left(-\frac{1}{2}\xi^T \Sigma \xi\right).$$

Ce qui conclut la preuve.

1.5.3 Loi du χ^2 et de Student

Lemme 22 (Origine de la loi du χ^2)**.**

Étant donné $\mathbf{X} = (X_1, \ldots, X_d)$ un vecteur gaussien de $\mathbb{R}^d$, d'espérance $m = \mathbb{E}[\mathbf{X}]$ et matrice de variance covariance $\mathbb{V}[\mathbf{X}] = \mathbb{I}_d$ alors la loi de la variable aléatoire réelle $\|\mathbf{X}\|_2^2 = X_1^2 + \ldots + X_d^2$ ne dépend que de d et de $|m|_2$.

Ce qui est dit dans ce théorème, c'est que la loi devrait dépendre de d et de m mais uniquement par $|m|_2$.

Preuve

Si $\|m\|_2 = 0$ alors $m = 0$ donc le résultat est trivial. Supposons que $\|m\|_2 > 0$ et prenons $\mathbf{Y}$ un vecteur gaussien de $\mathbb{R}^d$ suivant la loi $\mathcal{N}(m', \mathbb{I}_d)$ telle que $|m|_2 = |m'|_2$. Comme les normes sont égales, il existe une matrice U orthogonale telle que $m = Um'$ et, par linéarité (voir la proposition 15), nous avons $U\mathbf{Y} \sim \mathcal{N}\left(Um', UU^T\right) \overset{\mathcal{L}}{=} \mathcal{N}(m, \mathbb{I}_d)$ donc a la même loi que $\mathbf{X}$.
Or, nous avons que $\|\mathbf{Y}\|_2^2 = \|U\mathbf{Y}\|_2^2$ donc la loi de $\|\mathbf{Y}\|_2^2$ est la même que $\|U\mathbf{Y}\|_2^2$ qui la même que $\|\mathbf{X}\|_2^2$.

Définitions 22 (Loi du χ^2)**.**

Étant donné $\mathbf{X} = (X_1, \ldots, X_d)$ un vecteur gaussien de $\mathbb{R}^d$, d'espérance $m = \mathbb{E}[\mathbf{X}]$ et matrice de variance covariance $\mathbb{V}[\mathbf{X}] = \mathbb{I}_d$, la loi de $\|\mathbf{X}\|_2^2$ est appelé la **loi du χ^2 à d degrés de libertés et de paramètre de décentrage** $\|m\|_2^2$ et nous la notons $\chi^2\left(d, \|m\|_2^2\right)$. Lorsque $\|m\|_2^2 = 0$, nous parlons de **loi du χ^2 centrée à d degrés de libertés** et notons simplement $\chi^2(d)$.

Remarque

Lorsque nous parlons de *loi du khi deux* sans faire d'autres précisions, c'est la loi centrée qui est sous-entendue (de loin la plus utilisée).

Proposition 23 (Densité d'une loi du χ^2 centrée à d degrés de libertés)**.**

La loi du χ^2 centrée à d degrés de libertés a pour densité :

$$f(x) = \frac{1}{2^{d/2}\Gamma\left(\frac{d}{2}\right)} x^{\frac{d}{2}-1} e^{-\frac{x}{2}} \mathbb{1}_{\mathbb{R}_+^\star}(x)$$

où Γ est la fonction Γ définie sur $\mathbb{R}^+$ définie pour tout $x \in \mathbb{R}^+$ par :

$$\Gamma(x) = \int_0^{+\infty} t^{x-1} e^{-t} \mathrm{d}t.$$

Remarque

Cette loi est importante pour connaître la loi de l'estimateur de la variance.

Définitions 23 (Loi de Student)**.**

Étant données deux variables X et Y indépendantes de lois respectives $\mathcal{N}(\mu, 1)$ et $\chi^2(d)$, la loi de la variable

$$Z = \frac{X}{\sqrt{Y/d}}$$

est appelée **loi de Student à d degrés de liberté** que nous notons $\mathcal{T}(d, \mu)$. Si le **paramètre de décentrage** μ est nul alors nous notons simplement $\mathcal{T}(d)$.

Proposition 24 (Densité d'une loi de Student à d degrés de libertés)**.**

La loi de Student $\mathcal{T}(d)$ à d degrés de libertés a pour densité :

$$f(x) = \frac{1}{\sqrt{d\pi}} \frac{\Gamma\left(\frac{d+1}{2}\right)}{\Gamma\left(\frac{d}{2}\right)} \left(1 + \frac{x^2}{2}\right)^{-\frac{d+1}{2}}.$$

Remarque

Cette loi est importante pour connaître les intervalles de confiance de la variance.

Exercices 1.4

Cet exercice est tiré du livre Ycart et Genon-Catalot (2004). Étant donné $(X, Y, Z)^T$ un vecteur gaussien de moyenne $m = (1, -1, 1)^T$ et de matrice de variance covariance

$$K = \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix}.$$

Nous posons :

$$\begin{aligned} U &= -X + Y + Z, \\ V &= X - Y + Z, \\ W &= X + Y - Z. \end{aligned}$$

1. Quelles sont les lois de U, V et W ?

2. Déterminer l'espérance et la matrice de variance covariance du vecteur $(U, V, W)^T$. En déduire que les variables U, V et W sont indépendantes.
3. Utiliser ce qui précède pour écrire un algorithme de simulation de la loi $\mathcal{N}(m, K)$ si le logiciel ne permet de simuler que des gaussiennes univariées $\mathcal{N}(0, 1)$.
4. Quelle est la loi de $(U^2 + V^2 + W^2)/4$?

1.6 Indication pour réussir les exercices

Exercices 1.1

Le fait que ce soit une mesure vient de la proposition. Il reste juste à montrer que $\nu(F)$ est finie.

Exercices 1.2

Pour le premier point, il faut procéder par récurrence.
Pour le second, il faut se souvenir que l'intégrale sur $\mathbb{R}$ d'une fonction impaire est nulle (si elle est définie).

Exercices 1.3

La partie avec la fonction Γ est une constante positive donc ne doit pas vous perturber. Il s'agit simplement de dériver la fonction afin de trouver les maximums (s'ils existent).

Exercices 1.4

C'est une application directe du cours.

TABLE 1.1 – Table des probabilités $\mathbb{P}(X \leq x)$ où X suit une loi gaussienne centrée réduite et $x = d + c$.

d \ c	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08
0	0.5	0.504	0.508	0.512	0.516	0.5199	0.5239	0.5279	0.5319
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714
0.2	0.5793	0.5832	0.5871	0.591	0.5948	0.5987	0.6026	0.6064	0.6103
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.648
0.4	0.6554	0.6591	0.6628	0.6664	0.67	0.6736	0.6772	0.6808	0.6844
0.5	0.6915	0.695	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.719
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517
0.7	0.758	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823
0.8	0.7881	0.791	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.834	0.8365
1	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.877	0.879	0.881
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.898	0.8997
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306
1.5	0.9332	0.9345	0.9357	0.937	0.9382	0.9394	0.9406	0.9418	0.9429
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.975	0.9756	0.9761
2	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812
2.1	0.9821	0.9826	0.983	0.9834	0.9838	0.9842	0.9846	0.985	0.9854
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913
2.4	0.9918	0.992	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934
2.5	0.9938	0.994	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.996	0.9961	0.9962	0.9963
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.997	0.9971	0.9972	0.9973
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.998
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986
3	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.999
3.1	0.999	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.6	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.7	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.8	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999

Chapitre 2
Fondements de la statistique

"Le loto, c'est un impôt sur les gens qui ne comprennent pas les statistiques."

2.1 Objectifs

L'objectif de ce court chapitre est de comprendre les fondements de la statistique, la philosophie et le principe de modélisation.

2.2 Introduction

Commençons par quelques exemples d'application de la statistique :

Exemple

Adapté d'un exemple tiré du livre de Robert (2006). Considérant la proportion proportion de naissances masculines à Paris, Pierre-Simon de Laplace s'intéresse à savoir s'il y a plus de naissances d'hommes que de femmes. Pour cela, il va recenser 251 527 naissances masculines et 241 945 naissances féminines en 1785.
Une naissance peut être considérée comme le résultat d'une variable X_i de Bernoulli de paramètre $\theta \in [0; 1]$:

$$X_i = \begin{cases} \text{homme} & \text{avec probabilité } \theta, \\ \text{femme} & \text{avec probabilité } 1 - \theta. \end{cases}$$

$X_1, \ldots, X_n$ est alors l'échantillon des $n = 493\,472$ résultats de naissances observées. La question qu'on peut alors se poser est de savoir si $\theta = 1/2$, $\theta < 1/2$ ou $\theta > 1/2$.

Exemple

Adapté d'une histoire sur la *cryptographie automatique* de Ycart

(2017). Une ancienne technique de cryptographie consiste à associer chaque lettre de l'alphabet à une autre de façon unique : nous obtenons ainsi une bijection entre un texte lisible et un texte codé. Pour décoder, il suffit de faire la transformation inverse. Par exemple, avec le tableau de correspondance suivant :

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
K	R	D	U	C	W	M	G	B	E	F	P	N	H	I	Y	O	S	X	J	A	V	L	T	Q	Z

nous pouvons décrypter le texte :

KYYSCHUSC PK XJKJBXJBOAC CXJ SCYIXKHJ [1]

Une technique de *cryptanalyse* pourrait être d'essayer toutes les combinaisons possibles mais cela représente $26! \approx 4 \times 10^2 6$ possibilités pour la langue française (à la raison d'une possibilité toutes les secondes, cela prendrait un peu moins de 10^{10} fois l'âge de l'univers). Une autre idée a été proposée dès le $9^{\text{ème}}$ siècle par Al-Kindi basée sur une méthode statistique : en fait, il s'est aperçu que toutes les lettres ne revenaient pas de la même façon. Par exemple, dans la langue française, le e est celui qui revient le plus souvent (avec une fréquence de $12,10\%$ pour le corpus de wikipédia en 2008 par exemple si nous séparons les e sans accent de ceux avec accent [2]).
L'idée derrière est de dire que chaque lettre est le résultat d'une variable aléatoire X_i prenant ses valeurs dans $\{A, B, \ldots, Z\}$ avec une probabilité $(p_A, p_B, \ldots, p_Z)$. En calculant les fréquences dans le texte, nous pouvons essayer de retrouver le codage.
Notons toutefois que ce système demande des textes codés relativement longs. Pour accélérer l'estimation, il est conseillé de coupler avec les probabilités conditionnelles d'une lettre sachant une autre ; par exemple, la probabilité d'avoir deux w successifs est quasiment nulle.

En particulier, la procédure proposée par Alan Turing (1912-1954) pour décrypter Enigma fonctionna car les allemands envoyés régulièrement les bulletins météorologiques avec une chaîne de caractère régulière et assez longue à chaque fois (voir par exemple Lehning (2006)).

1. La solution étant **APPRENDRE LA STATISTIQUE EST REPOSANT** .

2. Voir par exemple `https://fr.wikipedia.org/wiki/Fr%C3%A9quence_d%27apparition_des_lettres_en_fran%C3%A7ais`

Exemple

Dans le film de présentation [3], nous nous étions intéressés à la hauteur d'un fleuve en fonction des pluies tombées dans les montagnes. Pour cela, nous supposons que chaque observation est le résultat d'une variable aléatoire Y_i qui dépend d'une co-variable x_i (la quantité de pluie tombée dans les montagnes) de façon polynomiale ; c'est-à-dire

qu'il existe un polynôme P_m de degrés m inconnu et de coefficients inconnus tels que pour tout $i \in \mathbb{N}$:

$$Y_i = P_m(x_i) + \varepsilon_i \text{ avec } \varepsilon_i \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, \sigma)$$

avec σ également inconnu. La question est alors d'estimer au mieux chacun des paramètres notamment le degré m du polynôme car la théorie de la sélection de modèle nous permet de voir que le polynôme qui collerait le mieux aux données serait un polynôme de degrés n (passant par toutes les observations) mais ce serait également le plus éloigné du vrai polynôme [4]. Ce phénomène est appelé du *surapprentissage*.

2.3 Modèle statistique

Commençons par quelques définitions.

Définitions 24 (Modèle statistique)**.**

Un **modèle statistique** est la donnée d'un espace mesurée $(E, \mathcal{E})$ et d'une famille $(\mathbb{P}_\theta)_{\theta \in \Theta}$ de mesures de probabilité d'indexation Θ. Le modèle associé est noté $(E, \mathcal{E}, \mathbb{P}_\theta, \theta \in \Theta)$.
Quand il existe $d \in \mathbb{N}^\star$ tel que $\Theta \subset \mathbb{R}^d$, nous disons que le modèle est **paramétrique**. Sinon, il est dit **non-paramétrique**.

Remarque

θ est inconnu mais son ensemble d'appartenance Θ est connu.

Exemple

Pour la loi de Bernoulli, nous savons que $\theta \in [0; 1]$.

⚠ Attention au piège

Le plus souvent, l'indexation se fait sur une seule famille de probabilités. Dans de très rares cas, nous pouvons imaginer une concurrence entre plusieurs familles de lois. Par exemple, sur l'ensemble fini $\mathbb{N}$, nous pouvons mettre en concurrence des lois de Poisson de paramètres $\lambda \in \mathbb{R}^+$ et des lois géométriques de paramètres $p \in [0; 1]$.

3. disponible ici : `https://www.youtube.com/watch?v=wQ-1Zj9sBlM`

4. Une vidéo pour illustrer ce problème est disponible ici : `https://www.youtube.com/watch?v=YuP1dOIbWfQ`

La démarche statistique consiste à donner de l'information sur le paramètre θ en s'appuyant sur la notion d'observation ; on parle alors d'**inférence** ou de **statistique inférentielle**.

Définition 25 (Observation)**.**

Une **observation** X est une variable aléatoire à valeur dans E et dont la loi appartient à la famille $(\mathbb{P}_\theta)_{\theta\in\Theta}$.

Point méthode (Hypothèse fondamentale)**.**

L'hypothèse fondamentale de la statistique est de supposer qu'une observation X suit une loi $\mathbb{P}_X$ et qu'il existe un paramètre $\theta \in \Theta$ tel que $\mathbb{P}_X = \mathbb{P}_\theta$.

Remarque

Dans le langage courant, nous confondons souvent l'observation X qui est une variable aléatoire et sa réalisation x qui est fixée.

Définition 26 (n-échantillon)**.**

Étant donné $n \in \mathbb{N}^\star$, un **n-échantillon de loi** $\mathbb{P}_X$ est la donnée de n variables aléatoires $X_1, \ldots, X_n$ indépendantes et identiquement distribuées selon la loi $\mathbb{P}$.

Remarque

Lorsque l'observation X a la structure d'un n-échantillon de loi $\mathbb{P}_\theta$ alors $\mathbb{P}_{\mathbf{X}} = (\mathbb{P}_\theta)^{\otimes n}$.

Exercices 2.1

Reprendre les exercices en début de cours et dire si les modèles sont paramétriques ou non-paramétriques. Dans le cas de modèles paramétriques, donnez l'espace des paramètres. Dans tous les cas, donner l'espace des observations.

Exemple

Une entreprise de téléphonie pense que le nombre d'appels journaliers passés par chaque client suit une loi de Poisson et souhaite tester cette hypothèse pour pouvoir fixer les tarifs. Pour ce faire, elle va choi-

sir aléatoirement des clients et récupérer le nombre d'appels. Formellement, les observations $X_1, \dots, X_n$ sont supposées indépendantes et de même loi et l'entreprise se demande donc si celle-ci suit une loi de Poisson et, dans ce cas, quel est le paramètre ?

Point méthode (Quelle loi choisir ?).

La problématique du choix de la loi dans le cas de modèles non-paramétriques peut être simplifiée en regardant l'ensemble de définition. Voici quelques exemples classiques :

- Si l'ensemble est dénombrable :
  - S'il n'a que deux éléments : loi de Bernoulli.
  - Si c'est $\{0, \dots, n\}$: loi binomiale, loi uniforme, loi hypergéométrique.
  - Si c'est $\mathbb{N}$ ou $\mathbb{N}^\star$: loi de Poisson, loi géométrique, loi binomiale négative.
- Si l'ensemble est non dénombrable :
  - Si c'est l'intervalle $[0, 1]$: loi uniforme, loi bêta.
  - Si c'est un intervalle borné : loi uniforme.
  - Si c'est $\mathbb{R}^+$: loi exponentielle, loi gamma, loi inverse gamma.
  - Si c'est $[a, +\infty[$ avec $a > 0$: loi de Pareto
  - Si c'est $\mathbb{R}$: loi gaussienne ou normale, loi de Cauchy, loi de Laplace.

Définition 27 (Identifiabilité).

Étant donné un modèle paramétré par θ, nous dirons que le modèle est **identifiable** si l'application $\theta \mapsto \mathbb{P}_\theta$ est injective.

Contre-exemple

Le contre-exemple le plus simple est celui de la loi gaussienne centrée de variance inconnue $\mathcal{N}(0, \theta^2)$:

- Si $\Theta = \mathbb{R}_+^\star$, le modèle est identifiable.
- Si $\Theta = \mathbb{R}^\star$, le modèle n'est pas identifiable car pour tout $\theta \neq 0$, les paramètres θ et $-\theta$ fournissent la même loi.

Exercices 2.2

Pour les lois mentionnées dans le point méthode précédent, donner l'espace de paramètre Θ de telle sorte que les lois correspondent à des modèles identifiables. Pour rappel, nous mettons les densités :

Nom	Abréviation	Support	Densité $f(x)$ ou $\mathbb{P}(X=x)$
Bernoulli	$\mathcal{B}(p)$	$\{0,1\}$	$p^x(1-p)^{1-x}$
Bêta	$\mathcal{B}e(a,b)$	$]0,1[$	$\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}x^{a-1}(1-x)^{b-1}$
Binomiale	$\mathcal{B}in(n;p)$	$\{0,1,\dots,n\}$	$\binom{n}{x}p^x(1-p)^{n-x}$
Binomiale négative	$\mathcal{NB}in(r;p)$	$\mathbb{N}$	$\binom{x+n-1}{x}p^n(1-p)^x$
Cauchy	$\mathcal{C}au(\mu,a)$	$\mathbb{R}$	$\frac{\pi}{1+\left(\frac{x-\mu}{a}\right)^2}$
Exponentielle	$\mathcal{E}(\lambda)$	$\mathbb{R}^+$	$\lambda e^{-\lambda x}$
Gamma	$\mathcal{G}a(\lambda)$	$\mathbb{R}_+^\star$	$\frac{\beta^\alpha}{\Gamma\alpha}x^{\alpha-1}e^{-\beta x}$
Gaussienne ou normale	$\mathcal{N}(\mu,\sigma^2)$	$\mathbb{R}$	$\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$
Géométrique	$\mathcal{G}éo(p)$	$\mathbb{N}^\star$	$(1-p)^{x-1}p$
Hypergéométrique	$\mathcal{H}(n,p,A)$	$\{0,1,\dots,n\}$	$\frac{\binom{pA}{x}\binom{(1-p)A}{n-x}}{\binom{A}{x}}$
Inverse gamma	$\mathcal{IG}a(\lambda)$	$\mathbb{R}_+^\star$	$\frac{\beta^\alpha}{\Gamma\alpha}x^{-\alpha-1}e^{\frac{-\beta}{x}}$
Laplace	Laplace(μ,b)	$\mathbb{R}$	$\frac{1}{2b}e^{-\frac{\lvert x-\mu\rvert}{b}}$
Pareto	$\mathcal{P}a(a,k)$	$[a,+\infty[$	$\frac{ka^k}{x^{k+1}}$
Poisson	$\mathcal{P}(\lambda)$	$\mathbb{N}$	$e^{-\lambda}\frac{\lambda^x}{x!}$
Uniforme	$\mathcal{U}(\{0,1,\dots,n\})$	$\{0,1,\dots,n\}$	$\frac{1}{n+1}$
Uniforme	$\mathcal{U}([a,b])$	$[a,b]$	$\frac{1}{b-a}$

Hypothèse.

Dans la suite de cours, nous supposerons généralement avoir un modèle statistique identifiable $(E,\mathcal{E},\mathbb{P}_\theta,\theta\in\Theta)$. De plus, nous noterons souvent $\mathbf{X}=(X_1,\dots,X_n)$ pour n-échantillon.

Chapitre 3
Estimation

3.1 Objectifs

Le but de ce chapitre est de comprendre le principe d'estimation et ce qui caractérise un *bon* estimateur. Nous allons donc introduire un certain nombre de notions et présenterons les estimateurs les plus connus.

3.2 Rappels/Pré-requis

Dans ce chapitre, nous aurons besoin des notions suivantes.

Probabilités : Différentes convergences, loi forte des grands nombres

3.3 Introduction

Le but de ce chapitre est d'estimer au mieux une quantité dépendante du paramètre inconnu θ. Pour bien différencier les outils que nous utilisons, nous allons commencer par un certain nombre de notations.

Définitions 28 (Paramètre d'intérêt)**.**

Étant donnée une famille de probabilité $(\mathbb{P}_\theta)_{\theta\in\Theta}$ et un n-échantillon $\mathbf{X} = (X_1, \dots, X_n)$, nous notons $\theta^\star$ le paramètre *inconnu* de la loi de X ; c'est-à-dire que $\mathbb{P}_X = \mathbb{P}_{\theta^\star}$.
La **quantité à estimer** ou **paramètre d'intérêt** est notée $g(\theta^\star)$ avec $g : \Theta \to \mathbb{R}^d$ où $d \in \mathbb{N}^\star$.

Remarque

Souvent, la quantité à estimer est simplement $\theta^\star$ c'est-à-dire pour g

valant l'identité.

Maintenant, nous allons parler de l'estimation :

Définition 29 (Estimateur)**.**

Étant donné un paramètre d'intérêt $g(\theta^\star)$, nous appelons un **estimateur** de $g(\theta^\star)$ toute application, notée $\hat{g}$, mesurable en l'observation X et indépendante de $\theta^\star$.

Remarque

Tout estimateur est donc de la forme $\hat{g} = h(X)$ avec h une application mesurable $E \to \mathbb{R}^d$.

Exemple

Par exemple, lorsque nous cherchons à estimer $\theta^\star$, nous proposons un estimateur noté $\widehat{\theta}$. Cet estimateur est donc une fonction des observations.

⚠ Attention au piège

La notion *d'indépendance de* $\theta^\star$ peut paraître étrange mais il arrive parfois que des auteurs proposent un estimateur directement lié à ce dernier. Par exemple, un estimateur de la forme $2\theta^\star$ ne fonctionne pas.

Remarque

On peut trouver toute sorte d'estimateurs. Un exemple naïf serait l'estimateur qui renvoie toujours la valeur 0 par exemple. Il faut donc réfléchir à obtenir un *bon* estimateur et en quel sens ?

3.4 Consistance d'un estimateur

La première propriété que nous voulons pour un estimateur est que son estimation soit relativement proche de la vraie valeur.

Définition 30 (Consistance d'un estimateur)**.**

Étant donné un n-échantillon $\mathbf{X} = (X_1, \ldots, X_n)$ et un estimateur $\hat{g}_n = h_n(X_1, \ldots, X_n)$, nous dirons que l'estimateur est

- **consistant** si pour tout $\theta^\star$ de Θ, $\hat{g}_n$ converge en $\mathbb{P}_{\theta^\star}$-probabilité vers $g(\theta^\star)$ lorsque n tend vers $+\infty$. C'est-à-dire que :
$$\forall \varepsilon > 0,\ \mathbb{P}\left(\|g(\theta^\star) - \hat{g}_n\|_{+\infty} > \varepsilon\right) \underset{n\to+\infty}{\longrightarrow} 0$$
où $\|\cdot\|_{+\infty}$ est la **norme infinie** sur $\mathbb{R}^d$ c'est-à-dire définie par :
$$\forall \mathbf{x} \in \mathbb{R}^d,\ \|x\|_{+\infty} = \max_{1\le j\le d} |x_j|.$$
- **fortement consistant** si pour tout $\theta^\star$ de Θ l'estimateur $\hat{g}_n$ converge $\mathbb{P}_{\theta^\star}$-presque sûrement vers $g(\theta^\star)$ lorsque n tend vers $+\infty$. C'est-à-dire que :
$$\mathbb{P}\left(\lim_{n\to+\infty} \hat{g}_n = g(\theta^\star)\right) = 1.$$

⚠ Attention au piège

En dimension finie, toutes les normes sont équivalentes donc nous pouvons choisir celle que nous préférons. En dimension infinie, il faut préciser par rapport à quelle norme nous avons la consistance.

Remarque

Par abus de langage, on parle souvent de consistance d'un estimateur là où on devrait plutôt parler de la consistance de la suite d'estimateurs.

Exemple

Dans l'exemple de la proportion de naissances masculines dans Paris, il est naturel d'estimer le paramètre $\theta^\star$ par la moyenne empirique
$$\hat{\theta}_n(X_1, \ldots, X_n) = \overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$
Alors, par la loi forte des grands nombres, nous savons que c'est un estimateur fortement consistant du paramètre.

D'autres propriétés peuvent être utilisées pour caractériser un bon estimateur, nous les verrons un peu plus tard.

3.5 Construction d'un estimateur

Dans cette section, nous présenterons deux types d'estimateurs : celui des moments et celui du maximum de vraisemblance.

3.5.1 Méthode des moments

Commençons par un exercice :

Exercices 3.1

Nous avons lancé un dé 1000 fois et nous avons obtenus 237 fois un chiffre inférieur ou égal à 3. Pensez-vous que le dé est équilibré ? Pourquoi ?

Modéliser l'expérience : comment d'écrire la situation ? Quelle loi est derrière ? Quel estimateur pourrions-nous proposer ? Quelle est sa loi exacte ? Sa loi asymptotique ? Que nous dit la loi forte des grands nombres ?

Le principe de la méthode des moments repose simplement sur la loi forte des grands nombres. Pour ce faire, nous supposons posséder un n-échantillon $\mathbf{X} = (X_1, \ldots, X_n)$ de loi $\mathbb{P}_{\theta^\star}$ et une fonction $\phi : E \to \mathbb{R}$ mesurable telle que $\phi(X_1)$ possède un moment d'ordre 1.
Pour estimer la valeur $g(\theta^\star) = \mathbb{E}_{\theta^\star}[\phi(X_1)]$, nous nous appuyons sur la moyenne empirique :

$$\widehat{g}_n = \frac{1}{n}\sum_{i=1}^{n} \phi(X_i)$$

qui, d'après la loi forte des grands nombres, converge presque sûrement vers $g(\theta^\star)$. Par conséquent, la suite d'estimateurs est fortement consistante.

Remarque

Cette méthode fonctionne dès que les paramètres d'intérêt peuvent s'exprimer comme des fonctions des moments de la loi des observations X_i.

Exemple

Nous estimons par exemple l'espérance $\mu_1 = \mathbb{E}[X]$ par la **moyenne empirique** $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Si les variables X_i sont à valeurs réelles et possèdent un moment d'ordre 2 alors nous pouvons estimer $\mu_2 = \mathbb{E}[X^2]$ par $\widehat{\mu}_{2,n} = \frac{1}{n}\sum_{i=1}^{n} X_i^2$.

Définition 31 (Méthode des moments (MM))**.**

Étant donnés un n-échantillon $(X_1, \ldots, X_n)$ de $\mathbb{R}^d$, q fonctions $\phi_1, \ldots, \phi_q$ allant respectivement de $\mathbb{R}^d$ dans $\mathbb{R}^{n_1}, \ldots, \mathbb{R}^{n_q}$ ayant chacune un moment d'ordre 1 ; c'est-à-dire que $\mathbb{E}_{\theta^\star}\left[\|\phi\left(X_1\right)\|\right]$ est finie et une fonction Ψ allant de $\prod_{j=1}^{q} \mathbb{R}^{n_j}$ dans $\mathbb{R}^m$, la **méthode des moments** ou **(MM)** consiste à estimer $g(\theta^\star) = \Psi\left(g_1\left(\theta^\star\right), \ldots, g_q\left(\theta^\star\right)\right)$ où $g_j(\theta^\star) = \mathbb{E}_{\theta^\star}\left[\phi_j(\theta^\star)\right]$ par

$$\widehat{g}_n = \Psi\left(\frac{1}{n}\sum_{i=1}^{n}\phi_1(X_i), \ldots, \frac{1}{n}\sum_{i=1}^{n}\phi_q(X_i)\right).$$

Exemple

Étant donné un n-échantillon $(X_1, \ldots, X_n)$ de loi admettant un moment d'ordre 2, la variance $\sigma^2 = \mathbb{V}_{\theta^\star}\left[X_1\right] = \mathbb{E}_{\theta^\star}\left[X_1^2\right] - \mathbb{E}_{\theta^\star}\left[X_1\right]^2$ est estimée par

$$\widehat{s}_n^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2$$

appelée la **variance empirique**.

Exemple

Étant donné un n-échantillon $(X_1, \ldots, X_n)$ de loi de Poisson $\mathcal{P}\left(\theta^\star\right)$ de paramètre $\theta^\star$ alors nous avons deux estimateurs des moments possibles pour $\theta^\star$:

- Comme $\mathbb{E}_{\theta^\star}\left[X_1\right] = \theta^\star$, nous pouvons prendre la moyenne empirique.
- Comme $\mathbb{V}_{\theta^\star}\left[X_1\right] = \theta^\star$, nous pouvons prendre la variance empirique.

Exemple

Étant donné un n-échantillon $(X_1, \ldots, X_n)$ de loi $\mathcal{N}\left(\mu, \sigma^2\right)$ alors :

- Si σ^2 est connu, la méthode des moments suggère d'estimer μ par la moyenne empirique.
- Si μ est connue, la méthode des moments suggère d'estimer σ^2 par

$$\widehat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \mu\right)^2.$$

- Si μ et σ^2 sont tous les deux inconnus alors la méthode des moments suggère d'estimer μ par la moyenne empirique et σ^2 par la

variance empirique.

3.5.2 Méthode du maximum de vraisemblance

⚠ Attention au piège

Dans cette partie, nous supposons que toutes les lois étudiées sont dominées presque sûrement par une mesure μ. Typiquement, nous aurons :

- la mesure de Lebesgue pour les lois sur $\mathbb{R}^d$.
- la mesure de comptage pour les cas de variables à valeurs dans un espace dénombrable.

Avant d'introduire la méthode, nous avons besoin d'une notation et d'une définition.

Notation

Pour tout paramètre $\theta \in \Theta$, nous notons f_θ la densité de $\mathbb{P}_\theta$ par rapport à la mesure dominante μ. En particulier, dans le cas discret, nous avons pour tout $x \in S$, $f_\theta(x) = \mathbb{P}_\theta(X = x) = \mathbb{P}(X = x; \theta)$.

Certains auteurs notent la densité par $p_\theta(\cdot)$ ou $p(\cdots; \theta)$.

Remarque

L'idée de la méthode du maximum de vraisemblance est de partir du constat que si nous avons une observation x et une famille de loi $(\mathbb{P}_\theta)_{\theta\in\Theta}$ alors nous pouvons calculer pour chaque θ la probabilité d'avoir eu l'observation. Par exemple, dans le cas discret, cela revient à calculer $\mathbb{P}_\theta(X = x) = \mathbb{P}(X = x; \theta)$. Une fois ces calculs faits, le θ le plus *vraisemblable* est simplement celui qui correspond à la plus forte probabilité d'avoir eu l'observation.

Définition 32 (Vraisemblance)**.**

Étant donné un n-échantillon $\mathbf{X} = (X_1, \ldots, X_n)$ de densité $f_{\theta^\star}$, nous appelons **vraisemblance du n-échantillon $\mathbf{X}$**, notée $V_{\mathbf{X}}$, la fonction :

$$\begin{aligned} V_{\mathbf{X}} : \Theta &\to \mathbb{R} \\ \theta &\mapsto V_{\mathbf{X}}(\theta^\star) = f_{\theta^\star}(\mathbf{X}) \end{aligned}$$

Définition 33 (Méthode du maximum de vraisemblance (MMV))**.**

Étant donné un n-échantillon $\mathbf{X} = (X_1, \ldots, X_n)$ de densité $f_{\theta^\star}$, la **méthode du maximum de vraisemblance (MMV)** consiste à estimer $g(\theta^\star) = \theta^\star$ par l'élément de Θ maximisant la vraisemblance $V_{\mathbf{X}}$.
Si un tel point $\widehat{\theta}_n$ existe, on l'appelle **l'estimateur du maximum de vraisemblance**.

⚠ Attention au piège

Nous pouvons faire trois remarques importantes :

- L'estimateur du maximum de vraisemblance n'existe pas toujours : c'est-à-dire qu'il n'y a pas toujours un point qui maximise la vraisemblance.
- S'il existe, il n'est pas obligatoirement unique : c'est-à-dire qu'il peut y avoir plusieurs points qui maximisent la vraisemblance.
- S'il existe et qu'il est unique, il n'est pas obligatoirement calculable : nous pouvons juste étudier la vraisemblance pour voir qu'elle doit admettre un maximum sans pouvoir trouver une formule explicite.

Proposition 25.

Dans le cas d'un n-échantillon $\mathbf{X} = (X_1, \ldots, X_n)$ de loi commune $\mathbb{P}_{\theta^\star}$, la vraisemblance prend la forme produit :

$$V_{\mathbf{X}}(\theta) = \prod_{i=1}^{n} f_\theta(X_i).$$

La conséquence est que nous préférons souvent maximiser la **log-vraisemblance** :

$$\log V_{\mathbf{X}}(\theta) = \sum_{i=1}^{n} \log f_\theta(X_i).$$

Preuve

Le produit vient de l'indépendance des variables $(X_i)_{i\in\{1,\ldots,n\}}$ et la formule du fait que nous avons une simple densité.

Remarque

L'inconvénient du maximum de vraisemblance est que nous n'estimons que le paramètre $\theta^\star$ mais pas des fonctions de celui-ci. En revanche, nous verrons par la suite que certains estimateurs du maximum de vraisemblance ont des meilleures propriétés que l'estimateur des moments.

Un calcul assez simple permet d'obtenir les estimateurs suivants :

Exercices 3.2

Dans le cas d'un n-échantillon $\mathbf{X} = (X_1, \ldots, X_n)$ de loi de Bernoulli $\mathcal{B}(\theta^\star)$ ou de Poisson $\mathcal{P}(\theta^\star)$, l'estimateur du maximum de vraisemblance est le même que celui des moments.

De même pour les modèles gaussiens $\mathcal{N}(\mu, \sigma^2)$ que nous connaissions μ ou σ^2 voir aucun des deux.

Un exemple assez intéressant d'estimateur du maximum de vraisemblance est le suivant :

Exemple

Étant donné un n-échantillon $\mathbf{X} = (X_1, \ldots, X_n)$ de même loi $\mathcal{U}([0, \theta^\star])$ avec $\theta^\star \in \mathbb{R}_+^\star$, nous avons :

$$\begin{aligned} V_{\mathbf{X}}(\theta) &= \prod_{i=1}^{n} \left(\frac{1}{\theta} \mathbb{1}_{[0,\theta]}(X_i) \right) \\ &= \frac{1}{\theta^n} \prod_{i=1}^{n} \mathbb{1}_{[0,\theta]}(X_i) \\ &= \theta^{-n} \mathbb{1}_{\left[X_{(n)}, +\infty\right[}(\theta) \text{ où } X_{(n)} = \max_{1 \leq i \leq n} X_i. \end{aligned}$$

Nous voyons que pour tout $n \in \mathbb{N}^\star$, la fonction $\theta \mapsto \theta^{-n}$ est décroissante et positive mais l'indicatrice $\theta \mapsto \mathbb{1}_{\left[X_{(n)}, +\infty\right[}(\theta)$ est nulle pour tout θ strictement plus petit que $X_{(n)}$. La valeur maximale de la fonction est donc atteinte pour $\widehat{\theta}_n = X_{(n)}$.

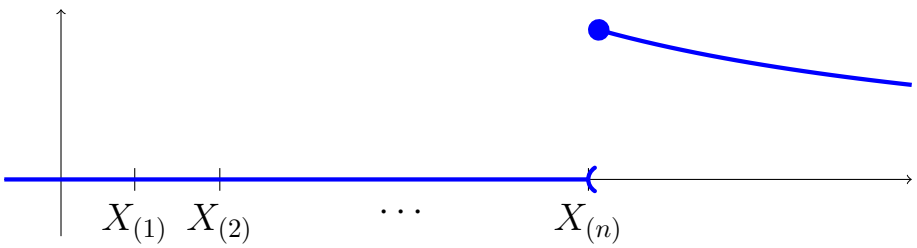


FIGURE 3.1 – Représentation en bleu de la fonction $\theta \mapsto \theta^{-n} \mathbb{1}_{\left[X_{(n)}, +\infty\right[}(\theta)$.

Enfin, voici un contre-exemple pour se rappeler qu'il n'est pas toujours possible de calculer le maximum de vraisemblance :

Exemple

Étant donné un n-échantillon $\mathbf{X} = (X_1, \ldots, X_n)$ de même loi $\mathcal{C}au\,(\theta^\star, 1)$ c'est-à-dire de densité

$$f_\theta(x) = \frac{1}{\pi} \times \frac{1}{1 + (x - \theta)^2}.$$

Le calcul de la log-vraisemblance est alors :

$$\log V_{\mathbf{X}}(\theta) = -n \log \pi - \sum_{i=1}^{n} \log \left[1 + (X_i - \theta)^2\right].$$

Cette fonction est définie sur $\mathbb{R}$, dérivable en θ, tend vers $-\infty$ lorsque θ tend vers $\pm\infty$ donc elle admet au moins un maximum qui est un point où la dérivée s'annule. En revanche, le calcul n'est pas explicite car la dérivée est une somme de fractions qui, une fois réduite au même dénominateur, est la somme de polynôme de degrés $2n - 1$. Il y a donc beaucoup de racines qui n'ont pas forcément de formules explicites (exceptées dans certains rares cas).

3.6 Qualité d'un estimateur

Dans cette partie, nous discutons des différentes propriétés que peut avoir un estimateur. Nous avons déjà vu l'indispensable consistance d'un estimateur qui permet de garantir que ce dernier sera *suffisamment proche* lorsque nous aurons suffisamment d'observations mais nous pouvons être intéressés par la vitesse à laquelle il va converger ou encore par la qualité non-asymptotique de son estimation.

3.6.1 Normalité asymptotique

Comme nous allons parler de convergence en loi, nous commençons par un petit rappel et un théorème que nous utiliserons.

Définition 34 (Convergence en loi)**.**

Étant donné une suite $(X_n)_{n\in\mathbb{N}^\star}$ de variables aléatoires d'un même espace de probabilité $(E, \mathcal{E}, \mathbb{P})$, nous disons que la suite $(X_n)_{n\in\mathbb{N}^\star}$ **converge en loi** vers la variable X si, pour toute fonction φ *continue bornée* sur E à valeurs dans $\mathbb{R}$, nous avons :

$$\lim_{n\to+\infty} \mathbb{E}\left[\varphi(X_n)\right] = \mathbb{E}\left[\varphi(X)\right].$$

Nous notons alors :

$$X_n \xrightarrow[n\to+\infty]{\mathcal{L}} X.$$

Remarque

Dans le cas de $\mathbb{R}^d$, nous pouvons montrer qu'il suffit de prendre des fonctions dans un espace dont l'adhérence englobe les fonctions continues à support compact sur $\mathbb{R}^d$ (voir par exemple le polycopié de Le Gall (2006), proposition 10.3.3). En particulier, nous pourrons être intéressés par prendre des fonctions lipschitzienne (voir la preuve du lemme 28 de Slutsky).

Dans notre cas, nous allons souvent utiliser la proposition suivante :

Proposition 26 (Cas de variables réelles)**.**

Étant donné une suite $(X_n)_{n\in\mathbb{N}^\star}$ de variables aléatoires d'un même espace de probabilité $(E, \mathcal{E}, \mathbb{P})$ et une variable X sur le même espace, nous avons l'équivalence des propositions suivantes :

(1) $(X_n)_{n\in\mathbb{N}^\star}$ converge en loi vers X ;

(2) Pour tout $x \in \mathbb{R}$ où F_X est continue (c'est-à-dire que $\mathbb{P}(X = x) = 0$), nous avons la convergence simple :

$$\lim_{n\to+\infty} F_{X_n}(x) = F_X(x).$$

Preuve

Une preuve complète peut être trouvée dans le polycopié de Le Gall (2006), proposition 10.3.2 et sa conséquence.

Dans un premier temps, nous allons nous concentrer sur l'estimation de $g(\theta^\star) = \theta^\star$. Le principe de normalité asymptotique va plus loin que la

consistance puisque nous nous intéressons à la vitesse de convergence de $\widehat{\theta}_n$ vers $\theta^\star$.

Définition 35 (Vitesse de convergence)**.**

Étant donné un estimateur $\widehat{g}_n$ du paramètre $g(\theta^\star)$ et une suite $(a_n)_{n\in\mathbb{N}^\star}$ tendant vers $+\infty$, nous disons que $(a_n)_{n\in\mathbb{N}^\star}$ est la **vitesse de convergence** de l'estimateur $\widehat{\theta}_n$ s'il existe une variable aléatoire Y de loi $\mathcal{L}$ non triviale indépendante de $\widehat{\theta}_n$ telle que

$$a_n\left(\widehat{\theta}_n - \theta^\star\right) \xrightarrow[n\to+\infty]{\mathcal{L}} Y.$$

Exemple

Étant donné un n-échantillon $\mathbf{X} = (X_1, \ldots, X_n)$ de loi commune admettant un moment d'ordre 2, d'espérance $\mu^\star$ et de matrice de variance covariance $\Sigma^\star$ alors, d'après le théorème de limite centrale, nous avons :

$$\sqrt{n}\left(\overline{X}_n - \mu^\star\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \Sigma^\star\right).$$

Sa vitesse de convergence est donc $\sqrt{n}$.

Exercices 3.3

Dans le cas d'un n-échantillon suivant une loi de Poisson $\mathcal{P}(\theta^\star)$, nous avons déjà qu'un estimateur possible est également $\widehat{\theta}_n = \overline{X}_n$ et nous aurons dans ce cas :

$$\sqrt{n}\left(\overline{X}_n - \theta^\star\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \theta^\star\right).$$

Certains auteurs utilisent parfois la notion d'équivalence asymptotique :

$$\overline{X}_n \underset{+\infty}{\overset{\mathcal{L}}{\sim}} \mathcal{N}\left(\theta^\star, \frac{\theta^\star}{n}\right).$$

Définition 36 (Asymptotiquement normal)**.**

Un estimateur $\widehat{g}_n$ est dit **asymptotiquement normal** s'il vérifie les trois conditions suivantes :

- La vitesse de convergence est en $\sqrt{n}$.
- La convergence a lieu en loi.
- La loi limite est une loi normale.

Remarque

Dans les cas classiques et notamment pour les estimateurs des moments, ce sera bien souvent le cas.

Contre-exemple

Si $X_1, \ldots, X_n$ est un n-échantillon suivant une loi $\mathcal{U}([0, \theta^\star])$ alors, nous avons vu que l'estimateur du maximum de vraisemblance était $\hat{\theta}_n = \max\{X_1, \ldots, X_n\} = X_{(n)}$ et calculons la loi de $n\left(\theta^\star - \hat{\theta}_n\right)$. Comme, $\mathbb{P}_{\theta^\star}$-presque sûrement, $\theta^\star$ est plus grand que toutes les variables X_i, nous avons que $n\left(\theta^\star - \hat{\theta}_n\right)$ est $\mathbb{P}_{\theta^\star}$-presque sûrement positif. Donc, pour tout $x \in \mathbb{R}^+$, nous avons :

$$
\begin{aligned}
F_{n\left(\theta^\star - \hat{\theta}_n\right)}(x) &= \mathbb{P}\left(n\left(\theta^\star - \hat{\theta}_n\right) \leq x\right) \\
&= \mathbb{P}\left(\theta^\star - \hat{\theta}_n \leq \frac{x}{n}\right) \\
&= \mathbb{P}\left(\theta^\star - \frac{x}{n} \leq \hat{\theta}_n\right) \\
&= 1 - \mathbb{P}\left(\theta^\star - \frac{x}{n} > \hat{\theta}_n\right) \\
&= 1 - \mathbb{P}\left(\max\{X_1, \ldots, X_n\} < \theta^\star - \frac{x}{n}\right) \\
&= 1 - \mathbb{P}\left(\left(X_1 < \theta^\star - \frac{x}{n}\right) \text{ et } \left(X_2 < \theta^\star - \frac{x}{n}\right) \text{ et } \ldots \text{ et } \left(X_n < \theta^\star\right.\right. \\
&= 1 - \prod_{i=1}^{n} \mathbb{P}\left(X_i < \theta^\star - \frac{x}{n}\right) \\
&= 1 - \prod_{i=1}^{n} \left(\frac{\theta^\star - \frac{x}{n}}{\theta^\star}\right) \\
&= 1 - \left(1 - \frac{x}{n\theta^\star}\right)^n \\
&\underset{n \to +\infty}{\longrightarrow} 1 - e^{-\frac{1}{\theta^\star}x}
\end{aligned}
$$

et nous reconnaissons la fonction de répartition de la loi exponentielle de paramètre $1/\theta^\star$. Par conséquent, la loi asymptotique de $n\left(\theta^\star - \hat{\theta}_n\right)$ est $\mathcal{E}\left(\frac{1}{\theta^\star}\right)$.

Remarque

Dans le contre-exemple précédent, nous voyons que la vitesse est en n par opposition à l'estimateur des moments qui converge à la vitesse $\sqrt{n}$.

Dans cette partie, la proposition la plus importante est celle de la méthode Delta :

Théorème 27 (Méthode Delta)**.**

Étant données une suite $(U_n)_{n\in\mathbb{N}^\star}$ de vecteurs aléatoires de $\mathbb{R}^d$ avec $d \in \mathbb{N}^\star$, une suite déterministe de réels $(a_n)_{n\in\mathbb{N}^\star}$ et une application $\ell : \mathbb{R}^d \to \mathbb{R}^p$ avec $p \in \mathbb{N}^\star$ telles que :

- a_n tend vers $+\infty$.
- Il existe $U^\star \in \mathbb{R}^d$ un vecteur déterministe et $V \in \mathbb{R}^d$ un vecteur aléatoire tels que

$$a_n \left(U_n - U^\star\right) \xrightarrow[n\to+\infty]{\mathcal{L}} V.$$

- ℓ est différentiable en $U^\star$ de différentielle notée $D\ell\left(U^\star\right) \in \mathcal{M}_{d\times p}\left(\mathbb{R}\right)$.

Alors, nous avons :

$$a_n \left(\ell(U_n) - \ell\left(U^\star\right)\right) \xrightarrow[n\to+\infty]{\mathcal{L}} D\ell\left(U^\star\right) V.$$

Pour la démonstration de ce théorème, nous avons besoin de rappeler le lemme de Slutsky.

Lemme 28 (Slutsky)**.**

Étant données $(X_n)_{n\in\mathbb{N}^\star}$ et $(Y_n)_{n\in\mathbb{N}^\star}$ deux suites de variables aléatoires prenant leurs valeurs respectivement dans $\mathbb{R}^d$ et $\mathbb{R}^p$ avec $d, p \in \mathbb{N}^\star$ éventuellement différents telles que X_n converge en loi vers un vecteur aléatoire X et Y_n converge en probabilité vers un vecteur déterministe C alors :

$$(X_n, Y_n) \xrightarrow[n\to+\infty]{\mathcal{L}} (X, C).$$

Remarque

Le plus souvent, nous appliquons à cette convergence jointe une fonction g ce qui permet de déduire :

$$g\left(X_n, Y_n\right) \xrightarrow[n\to+\infty]{\mathcal{L}} g(X, C).$$

Les fonctions les plus utilisées sont par exemple la somme (si $d = p$), le produit (si $p = 1$) ou la division (si $C \neq 0$ et $p = 1$).

Preuve

Pour démontrer le lemme de Slutsky, nous prenons une fonction $\varphi : \mathbb{R}^d \times \mathbb{R}^p \to \mathbb{R}$ lipshitzienne et bornée de constante de lipschitz L et M un majorant de la fonction. Commençons par utiliser l'inégalité triangulaire pour montrer que :

$$|\mathbb{E}[\varphi(X_n, Y_n)] - \mathbb{E}[\varphi(X, C)]| \leq \underbrace{|\mathbb{E}[\varphi(X_n, Y_n)] - \mathbb{E}[\varphi(X_n, C)]|}_{=(1)} + \underbrace{|\mathbb{E}[\varphi(X_n, C)] - \mathbb{E}[}_{=(2)}$$

Comme la fonction $x \mapsto \varphi(x, C)$ est bornée et que la suite $(X_n)_{n\in\mathbb{N}^\star}$ tend en loi vers X alors, par le théorème du porte-manteau, le second terme (2) tend vers 0.

Pour le premier terme (1), nous prenons $\varepsilon > 0$ et nous utilisons le fait que la fonction soit lipshitzienne :

$$\begin{aligned}
(1) &= |\mathbb{E}[\varphi(X_n, Y_n)] - \mathbb{E}[\varphi(X_n, C)]| \\
&\leq \mathbb{E}[|\varphi(X_n, Y_n) - \varphi(X_n, C)|] \\
&\leq \mathbb{E}\left[|\varphi(X_n, Y_n) - \varphi(X_n, C)\| \,\middle|\, \|Y_n - C\| > \varepsilon\right] \mathbb{P}(\|Y_n - C\| > \varepsilon) \\
&\quad + \mathbb{E}\left[|\varphi(X_n, Y_n) - \varphi(X_n, C)\| \,\middle|\, \|Y_n - C\| \leq \varepsilon\right] \mathbb{P}(\|Y_n - C\| \leq \varepsilon) \\
&\leq \mathbb{E}\left[|\varphi(X_n, Y_n)| + |\varphi(X_n, C)\| \,\middle|\, \|Y_n - C\| > \varepsilon\right] \mathbb{P}(\|Y_n - C\| > \varepsilon) \\
&\quad + \mathbb{E}\left[L\|Y_n - C\| \,\middle|\, \|Y_n - C\| \leq \varepsilon\right] \mathbb{P}(\|Y_n - C\| \leq \varepsilon) \\
&\leq \mathbb{E}\left[2M \middle| \|Y_n - C\| > \varepsilon\right] \mathbb{P}(\|Y_n - C\| > \varepsilon) \\
&\quad + \mathbb{E}\left[L\varepsilon \middle| \|Y_n - C\| \leq \varepsilon\right] \mathbb{P}(\|Y_n - C\| \leq \varepsilon) \\
&\leq 2M\mathbb{P}(\|Y_n - C\| > \varepsilon) + L\varepsilon\mathbb{P}(\|Y_n - C\| \leq \varepsilon)
\end{aligned}$$

Comme $(Y_n)_{n\in\mathbb{N}^\star}$ converge en probabilité vers C, nous avons pour tout $\varepsilon > 0$:

$$\limsup_{n\to+\infty} |\mathbb{E}[\varphi(X_n, Y_n)] - \mathbb{E}[\varphi(X_n, C)]| \leq L\varepsilon$$

ce qui permet de conclure.

À l'aide de ce lemme, nous pouvons maintenant démontrer la méthode Delta :

Preuve

Avant de commencer, rappelons que la définition de la différentiabilité est la suivante :

$$\forall M > 0,\ \forall \varepsilon > 0,\ \exists \delta > 0,$$
$$\forall U' \in \mathbb{R}^d, \|U' - U^\star\| \leq \delta \Rightarrow \|\ell(U') - \ell(U^\star) - D\ell(U^\star)(U' - U^\star)\| \leq \frac{M}{\varepsilon}\|U' - U^\star\|$$

Commençons par décomposer la variable étudiée pour introduire la notion de différentiabilité :

$$a_n(\ell(U_n) - \ell(U^\star)) = \underbrace{a_n(\ell(U_n) - \ell(U^\star)) - D\ell(U^\star)\, a_n(U_n - U^\star)}_{=(1)} + \underbrace{D\ell(U^\star)\, a_n(U_n -}_{=(2)}$$

et montrons que le premier terme converge vers 0 et en probabilité. Pour cela, nous fixons $\varepsilon > 0$ et $M > 0$ et nous choisissons $\delta > 0$ tel que

$$\begin{aligned}
\mathbb{P}(\|(1)\| > \varepsilon) &= \mathbb{P}(\|a_n(\ell(U_n) - \ell(U^\star)) - D\ell(U^\star)\, a_n(U_n - U^\star)\| > \varepsilon) \\
&= \mathbb{P}(a_n\|\ell(U_n) - \ell(U^\star) - D\ell(U^\star)(U_n - U^\star)\| > \varepsilon) \\
&= \mathbb{P}(a_n\|\ell(U_n) - \ell(U^\star) - D\ell(U^\star)(U_n - U^\star)\| > \varepsilon | \|U_n - U^\star\| \leq \delta) \\
&\quad + \mathbb{P}(a_n\|\ell(U_n) - \ell(U^\star) - D\ell(U^\star)(U_n - U^\star)\| > \varepsilon | \|U_n - U^\star\| > \\
&\leq \mathbb{P}(a_n\|U_n - U^\star\| > M | \|U_n - U^\star\| \leq \delta) + \mathbb{P}(a_n\|U_n - U^\star\| > a_n \\
&\leq \mathbb{P}(a_n\|U_n - U^\star\| > M) + \mathbb{P}(a_n\|U_n - U^\star\| > a_n\delta) \\
&\leq 2\mathbb{P}(a_n\|U_n - U^\star\| > \min(M, a_n\delta)).
\end{aligned}$$

Ceci étant vrai pour tout $n \in \mathbb{N}^\star$ et comme a_n tend vers l'infini, nous avons pour n assez grand :

$$\begin{aligned}
\limsup_{n\to+\infty} \mathbb{P}(\|(1)\| > \varepsilon) &\leq \limsup_{n\to+\infty} 2\mathbb{P}(a_n\|U_n - U^\star\| > M)) \\
&\leq 2\mathbb{P}(\|V\| \geq M)
\end{aligned}$$

car $a_n(U_n - U^\star)$ converge en loi vers V et par le lemme de Porte-Manteau. Or, ceci est vrai pour tout M donc nous pouvons faire tendre M vers $+\infty$ et nous avons donc la convergence en probabilité vers 0.

Enfin, par les hypothèses, nous savons que la deuxième partie (2) converge en loi vers $D\ell(U^\star)V$ donc, en utilisant le lemme de Slutsky sur la somme, nous avons le résultat.

Exemple

Étant donné un n-échantillon $\mathbf{X} = (X_1, \ldots, X_n)$ suivant une loi de Poisson $\mathcal{P}(\theta^\star)$, nous avons vu qu'un estimateur possible de $\theta^\star$ est

la variance empirique $\hat{\theta}_{n,2} = \overline{X_n^2} - \overline{X}_n^2$. Par la méthode Delta, nous montrons que :

$$\sqrt{n}\left(\hat{\theta}_{n,2} - \theta^\star\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, 2\theta^{\star 2} + \theta^\star\right).$$

Exercices 3.4

Montrer le résultat de l'exemple.

Corollaire 29 (Méthode Delta et estimateur des moments)**.**

Étant donnés un n-échantillon $\mathbf{X} = (X_1, \ldots, X_n)$ de vecteurs de $\mathbb{R}^d$, de loi commune admettant un moment d'ordre 2, d'espérance μ et de matrice de variance-covariance Σ et ℓ une fonction différentiable en μ de différentielle notée $D\ell(\mu)$ alors

$$\sqrt{n}\left(\ell\left(\overline{X}_n\right) - \ell(\mu)\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, D\ell(\mu)\Sigma D\ell(\mu)^T\right).$$

3.6.2 Estimation sans biais

Le principe d'estimation est de vérifier si on peut *espérer* avoir une estimation parfaite en moyenne.

Définitions 37 (Biais)**.**

Étant donnée une observation X et un estimateur $\hat{g} = h(X)$ de $g(\theta^\star)$ ayant un moment d'ordre 1 ; c'est-à-dire tel que $\mathbb{E}_{\theta^\star}[\|\hat{g}\|] < +\infty$, pour tout $\theta^\star \in \Theta$, nous appelons **biais** la fonction $b : \Theta \to \mathbb{R}^d$ définie pour tout $\theta^\star \in \Theta$ par :

$$b(\theta^\star) = \mathbb{E}_{\theta^\star}[\hat{g}] - g(\theta^\star).$$

L'estimateur $\hat{g}$ sera dit **sans biais** si $b(\theta^\star) = 0$ pour tout $\theta^\star \in \Theta$; ou encore si l'espérance $\mathbb{E}_{\theta^\star}[\hat{g}] = g(\theta^\star)$.

Exemple

Lorsqu'un moment d'ordre 1 existe, la moyenne empirique est toujours un estimateur sans biais ; en revanche, celui de la variance est biaisé.

Astuce algorithmique.

Généralement, nous trouvons l'estimateur non biaisé de la variance dans les logiciels par défaut. De plus en plus, nous avons aussi l'estimateur classique.

Exemple

Dans le cas d'un n-échantillon de loi $\mathcal{U}([0, \theta^\star])$, l'estimateur des moments est sans biais mais celui du maximum de vraisemblance est biaisé (il sous-estime la vraie valeur). Néanmoins, nous préférons le second car il converge plus vite vers la bonne valeur.

Contre-exemple

Dans le cas d'un modèle $\mathcal{N}(\theta^\star, 1)$ où $\theta^\star \in \mathbb{R}$, l'estimateur $\hat{g} = 0$ est nul pour $\theta^\star = 0$ mais différent de zéro dès que $\theta^\star \neq 0$, il est donc biaisé puisqu'il ne vérifie pas la condition pour tout $\theta^\star \in \mathbb{R}$.

Remarque

Nous pouvons faire trois remarques :

1. Le caractère non-biaisé d'un estimateur doit être vérifier pour tout $\theta^\star \in \Theta$; c'est-à-dire que, quelque soit la quantité à estimer, le biais doit être nul (pas juste pour des valeurs).
2. Le critère est non-asymptotique.
3. Le critère peut être discutable : à un moment, le biais était jugé très important (c'est pour cela que beaucoup de logiciels proposent des versions non biaisées par défaut de certains estimateurs). Néanmoins, il vaut toujours mieux un estimateur qui converge plus vite vers la bonne valeur qu'un estimateur non biaisé. De même, un estimateur non biaisé qui possède une forte variance a parfois une plus forte probabilité de fournir une estimation très loin de la quantité à estimer qu'un paramètre biaisé avec une faible variance.

Par rapport à la troisième remarque, nous préférons nous contenter de la version asymptotique.

Définition 38 (Asymptotiquement sans biais)**.**

Étant donnée une observation X et un estimateur $\hat{g} = h(X)$ de $g(\theta^\star)$ ayant un moment d'ordre 1 ; c'est-à-dire tel que $\mathbb{E}_{\theta^\star}[\|\hat{g}\|] < +\infty$, pour tout $\theta^\star \in \Theta$, l'estimateur $\hat{g}_n$ sera dit **asymptotiquement sans biais** si pour tout $\theta^\star \in \Theta$:

$$\lim_{n \to +\infty} \mathbb{E}_{\theta^\star}[\hat{g}_n] = g(\theta^\star).$$

Point méthode.

Pour le biais, il suffit donc de procéder en deux étapes :

1. Nous calculons $\mathbb{E}_{\theta^\star}[\hat{g}_n]$:
   - S'il vaut $g(\theta^\star)$ pour tout $n \in \mathbb{N}^\star$, l'estimateur est sans biais donc asymptotiquement sans biais aussi.
2. Sinon, nous regardons la limite :
   - Si elle vaut $g(\theta^\star)$, l'estimateur est asymptotiquement sans biais aussi.
   - Sinon, il peut y avoir des valeurs de $\theta^\star \in \Theta$ pour lesquelles, l'estimateur n'est pas consistant.

3.6.3 Estimation optimale

Dans cette partie, nous allons parler du risque d'un estimateur qui représente à quel point un estimateur peut s'écarter ou non de la valeur qu'il estime et nous essayerons de minimiser ce risque.

Risque d'un estimateur

Commençons par introduire le risque.

Définition 39 (Risque d'un estimateur)**.**

Étant donné un estimateur $\hat{g}$ de $g(\theta^\star)$, nous associons son **risque (quadratique)** défini comme l'application suivante :

$$\begin{array}{rcl} R : \Theta & \to & [0, +\infty] \\ \theta^\star & \mapsto & \mathbb{E}_{\theta^\star}\left[(\hat{g} - g(\theta^\star))^2\right] \end{array}$$

Remarque

Si plusieurs estimateurs sont étudiés en même temps, il est préférable d'indexer la fonction R par le nom de l'estimateur.

Un estimateur sera d'autant meilleur que son risque est faible sur Θ.

Proposition 30.

Étant donné un estimateur $\hat{g}$ de $g(\theta^\star)$, nous avons la relation suivante :

$$R(\theta^\star) = \left(\mathbb{E}_{\theta^\star}[\hat{g}] - g(\theta^\star)\right)^2 + \mathbb{E}_{\theta^\star}\left[(\hat{g} - \mathbb{E}_{\theta^\star}[\hat{g}])^2\right] = b(\theta^\star)^2 + \mathbb{V}_{\theta^\star}[\hat{g}].$$

Preuve

La démonstration consiste simplement à introduire l'espérance :

$$\begin{aligned} R(\theta^\star) &= \mathbb{E}_{\theta^\star}\left[(\hat{g} - g(\theta^\star))^2\right] \\ &= \mathbb{E}_{\theta^\star}\left[(\hat{g} - \mathbb{E}_{\theta^\star}[\hat{g}] + \mathbb{E}_{\theta^\star}[\hat{g}] - g(\theta^\star))^2\right] \\ &= \mathbb{E}_{\theta^\star}\left[(\hat{g} - \mathbb{E}_{\theta^\star}[\hat{g}])^2\right] + 2\mathbb{E}_{\theta^\star}\left[(\hat{g} - \mathbb{E}_{\theta^\star}[\hat{g}])\left(\mathbb{E}_{\theta^\star}[\hat{g}] - g(\theta^\star)\right)\right] + \mathbb{E}_{\theta^\star}\left[(\mathbb{E}_{\theta^\star}[\hat{g}] - \right. \\ &= \mathbb{V}_{\theta^\star}[\hat{g}] + 2\mathbb{E}_{\theta^\star}\left[\hat{g} - \mathbb{E}_{\theta^\star}[\hat{g}]\right]\left(\mathbb{E}_{\theta^\star}[\hat{g}] - g(\theta^\star)\right) + b(\theta^\star)^2 \\ &= \mathbb{V}_{\theta^\star}[\hat{g}] + 2\underbrace{\left(\mathbb{E}_{\theta^\star}[\hat{g}] - \mathbb{E}_{\theta^\star}[\hat{g}]\right)}_{=0}\left(\mathbb{E}_{\theta^\star}[\hat{g}] - g(\theta^\star)\right) + b(\theta^\star)^2 \\ &= b(\theta^\star)^2 + \mathbb{V}_{\theta^\star}[\hat{g}] \end{aligned}$$

Corollaire 31.

Étant donné un estimateur $\hat{g}$ sans biais de $g(\theta^\star)$, le risque est égal à la variance de l'estimateur.

Dans le cadre des estimateurs non biaisés, nous nous intéressons donc à ceux qui minimisent uniformément la variance.

Définition 40.

Estimateur UMVU Étant donné une collection $\mathcal{G}$ d'estimateurs non biaisés de $g(\theta^\star)$, c'est-à-dire un ensemble d'applications mesurable en l'observation X à valeurs dans l'espace d'arrivée de $g(\theta^\star)$, nous appelons **estimateur UMVU** pour (**Uniform Minimum Variance Unbiased**) tout estimateur $\widehat{g}$ qui minimise la variance de façon uniforme c'est-à-dire vérifiant :

$$\forall \theta \in \Theta, \forall \widehat{h} \in \mathcal{G}, \mathbb{V}_{\theta^\star}\left[\widehat{g}\right] \leq \mathbb{V}_{\theta^\star}\left[\widehat{h}\right].$$

Information de Fisher

Dans cette partie et les suivantes, nous avons besoin de restreindre l'ensemble des lois étudiées.

Définition 41 (Modèle paramétrique régulier)**.**

Étant donné un modèle paramétrique $(f_\theta)_{\theta\in\Theta}$, nous dirons qu'il est **régulier** s'il vérifie les trois conditions suivantes :

1. Θ est un ouvert de $\mathbb{R}^p$.
2. Le support des densités f_θ, c'est-à-dire l'ensemble $\left\{x \in \mathbb{R}^d \middle| f_\theta(x) > 0\right\}$, est indépendant de θ et la fonction $\theta \mapsto f_\theta$ est dérivable sur Θ sauf éventuellement pour un nombre fini de points.
3. Pour toute application h mesurable en X dont l'espérance du module est finie sous la loi f_θ pour tout θ, c'est-à-dire que $\mathbb{E}_\theta\left[\|h(X)\|\right] < +\infty$, nous supposons que l'application $\theta \mapsto \int_{\mathcal{X}} g(x) f_\theta(x)\mathrm{d}x$ est dérivable et que sa dérivée vaut :

$$\frac{\partial}{\mathrm{d}\theta}\int_{\mathcal{X}} g(x) f_\theta(x)\mathrm{d}x = \int_{\mathcal{X}} g(x)\frac{\partial}{\mathrm{d}\theta} f_\theta(x)\mathrm{d}x$$

où $\frac{\partial}{\mathrm{d}\theta}$ représente le gradient (donc si $p = 1$, c'est juste la dérivée).

Remarque

Pour les deux premiers points, nous pouvons faire des remarques :

1. Un exemple classique d'ouverts de $\mathbb{R}^p$ est le pavé $]a_1, b_1[\times]a_2, b_2[\times\cdots\times]a_p, b_p[$ avec $(a_1, b_1, a_2, \ldots, b_p) \in \overline{\mathbb{R}}^{2p}$.
2. La loi uniforme $\mathcal{U}([0, \theta])$ ne vérifie pas ces conditions. Ceci ne veut pas dire que les résultats que nous démontrerons par la suite ne s'appliquent pas forcément ; par contre, les démonstrations seront plus compliquées.

Définition 42 (Score)**.**

Étant donné un modèle paramétrique régulier $(f_\theta)_{\theta\in\Theta}$ et un n-échantillon $\mathbf{X} = (X_1, \ldots, X_n)$, nous appelons **score**, noté $U(\cdot)$, la fonction $U : \theta \mapsto \mathbb{R}$ définie par

$$U(\theta) = \frac{\partial}{\mathrm{d}\theta} \log V_{\mathbf{X}}(\theta) = \frac{\partial}{\mathrm{d}\theta} \left(\sum_{i=1}^{n} \log f_\theta(X_i) \right).$$

Propriétés 32.

Dans un modèle paramétrique régulier, nous avons pour tout $\theta \in \Theta$:

$$\mathbb{E}_\theta \left[U(\theta) \right] = 0.$$

Preuve

Pour montrer cela, il suffit de faire le calcul. Dans le cas continu, nous avons :

$$\begin{aligned}
\mathbb{E}_\theta \left[U(\theta) \right] &= \mathbb{E}_\theta \left[\frac{\partial}{\mathrm{d}\theta} \left(\sum_{i=1}^{n} \log f_\theta(X_i) \right) \right] \\
&= \sum_{i=1}^{n} \mathbb{E}_\theta \left[\frac{\partial}{\mathrm{d}\theta} \log f_\theta(X_i) \right] \\
&= n \int_{\mathcal{X}} \frac{\partial}{\mathrm{d}\theta} \log f_\theta(x) f_\theta(x) \mathrm{d}x \\
&= \sum_{i=1}^{n} \int_{\mathcal{X}} \frac{\frac{\partial}{\mathrm{d}\theta} f_\theta(x)}{f_\theta(x)} f_\theta(x) \mathrm{d}x \\
&= \sum_{i=1}^{n} \int_{\mathcal{X}} \frac{\partial}{\mathrm{d}\theta} f_\theta(x) \mathrm{d}x
\end{aligned}$$

$$= \sum_{i=1}^{n} \frac{\partial}{\mathrm{d}\theta} \underbrace{\int_{\mathcal{X}} f_\theta(x)\mathrm{d}x}_{=1}$$
$$= 0.$$

Remarque

L'estimateur du maximum de vraisemblance peut être vu comme la valeur annulant le score.

Définition 43 (Information de Fisher)**.**

Étant donné un modèle paramétrique régulier $(f_\theta)_{\theta\in\Theta}$ et un n-échantillon $\mathbf{X} = (X_1, \ldots, X_n)$, nous appelons **information de Fisher**, notée $I_n(\cdot)$, la matrice symétrique de taille $p \times p$ définie pour tout $\theta \in \Theta$ par :

$$I_n(\theta) = \mathbb{E}_{\theta^\star}\left[U(\theta)U(\theta)^T\right] = \left(\mathbb{E}\left[\frac{\partial \log V_{\mathbf{X}}(\theta)}{\mathrm{d}\theta_i}\frac{\partial \log V_{\mathbf{X}}(\theta)}{\mathrm{d}\theta_j}\right]\right)_{1\leq i,j\leq p}.$$

Bornes de Cramer-Rao

Dans cette partie, nous faisons des hypothèses supplémentaires sur l'ensemble de définitions et sur la régularité des fonctions :

Hypothèse.

Nous supposons que nous sommes dans un modèle paramétrique régulier $(f_\theta)_{\theta\in\Theta}$ et que $\Theta \subset \mathbb{R}$.

Corollaire 33 (Cas unidimensionnel)**.**

Dans le cas où $\theta \in \mathbb{R}$, nous avons :

$$I_n(\theta) = nI_1(\theta) = n\int \frac{\left[\frac{\partial}{\mathrm{d}\theta} f_\theta(x)\right]^2}{f_\theta(x)} \mathbb{1}_{\{f_\theta(x)>0\}}\mathrm{d}x.$$

Preuve

Le fait que $I_n(\theta) = I_1(\theta)$ vient de l'indépendance des observations

et la deuxième partie est simplement le calcul.

Remarque

Ce corolaire permet de ne se concentrer que sur les propriétés de $I_1(\theta)$ que nous trouvons souvent simplement sous la notation $I(\theta)$.

Corollaire 34 (Information de Fisher et variance)**.**

Étant donné un modèle paramétrique régulier $(f_\theta)_{\theta\in\Theta}$ unidimensionnel, nous avons :

$$I_1(\theta) = \mathbb{V}_\theta\left[\frac{\partial}{\mathrm{d}\theta}\log f_\theta(X)\right].$$

Preuve

Pour cela, nous faisons le calcul de la variance :

$$\begin{aligned}\mathbb{V}_\theta\left[\frac{\partial}{\mathrm{d}\theta}\log f_\theta(X)\right] &= \mathbb{E}_\theta\left[\left(\frac{\partial}{\mathrm{d}\theta}\log f_\theta(X)\right)^2\right] - \mathbb{E}_\theta\left[\frac{\partial}{\mathrm{d}\theta}\log f_\theta(X)\right]^2 \\ &= I_1(\theta) - \underbrace{\mathbb{E}_\theta\left[U(\theta)\right]^2}_{=0}.\end{aligned}$$

Proposition 35 (Information de Fisher et dérivée seconde)**.**

Étant donné un modèle paramétrique régulier $(f_\theta)_{\theta\in\Theta}$ unidimensionnel, si nous pouvons dériver deux fois sous le signe somme alors nous avons :

$$I_1(\theta) = -\mathbb{E}_\theta\left[\frac{\partial^2}{\mathrm{d}\theta^2}\log f_\theta(X)\right].$$

Preuve

Il suffit de faire le calcul :

$$\begin{aligned}\mathbb{E}_\theta\left[\frac{\partial^2}{\mathrm{d}\theta^2}\log f_\theta(X)\right] &= \int_{\{x\in\mathcal{X}|f_\theta(x)>0\}}\left[\frac{\partial^2}{\mathrm{d}\theta^2}\log f_\theta(x)\right] f_\theta(x)\mathrm{d}x \\ &= \int_{\{x\in\mathcal{X}|f_\theta(x)>0\}}\left[\frac{\partial}{\mathrm{d}\theta}\frac{\frac{\partial}{\mathrm{d}\theta}f_\theta(x)}{f_\theta(x)}\right] f_\theta(x)\mathrm{d}x\end{aligned}$$

$$
\begin{aligned}
&= \int_{\{x\in\mathcal{X}|f_\theta(x)>0\}} \left[\frac{\frac{\partial^2}{\mathrm{d}\theta^2}f_\theta(x)}{f_\theta(x)} - \frac{\left[\frac{\partial}{\mathrm{d}\theta}f_\theta(x)\right]^2}{f_\theta^2(x)}\right] f_\theta(x)\mathrm{d}x \\
&= \int_{\{x\in\mathcal{X}|f_\theta(x)>0\}} \frac{\partial^2}{\mathrm{d}\theta^2}f_\theta(x)\mathrm{d}x - \int_{\{x\in\mathcal{X}|f_\theta(x)>0\}} \frac{\left[\frac{\partial}{\mathrm{d}\theta}f_\theta(x)\right]^2}{f_\theta(x)}\mathrm{d}x \\
&= \frac{\partial^2}{\mathrm{d}\theta^2}\int_{\{x\in\mathcal{X}|f_\theta(x)>0\}} f_\theta(x)\mathrm{d}x - I_1(\theta) \\
&= -I_1(\theta)
\end{aligned}
$$

Exercices 3.5

Montrer de deux façons différentes que pour un n-échantillon de loi de Poisson, nous avons $I_n(\theta) = n/\theta$.

Théorème 36 (Borne de Cramer-Rao)**.**

Étant donné un estimateur $\widehat{g} = h(X)$ sans biais de $g(\theta)$ avec g une fonction dérivable sur Θ, si le modèle est régulier d'information de Fisher strictement positive sur Θ et si $\mathbb{E}_\theta\left[h^2(X)\right]$ est bornée alors nous avons :

$$\forall\theta\in\Theta,\ \mathbb{V}_\theta\left[\widehat{g}\right] \geq \frac{\left(g'(\theta)\right)^2}{I_1(\theta)}.$$

Preuve

Comme $\hat{g}$ est supposé sans biais, nous savons que $\mathbb{E}_\theta[\hat{g}] = g(\theta)$ donc nous avons :

$$\begin{aligned}
g'(\theta) &= \frac{\partial}{\mathrm{d}\theta} g(\theta) \\
&= \frac{\partial}{\mathrm{d}\theta} \mathbb{E}_\theta[\hat{g}] \\
&= \frac{\partial}{\mathrm{d}\theta} \mathbb{E}_\theta[h(X)] \\
&= \frac{\partial}{\mathrm{d}\theta} \int_{\{x \in \mathcal{X} | f_\theta(x) > 0\}} h(x) f_\theta(x) \mathrm{d}x \\
&= \int_{\{x \in \mathcal{X} | f_\theta(x) > 0\}} h(x) \frac{\partial}{\mathrm{d}\theta} f_\theta(x) \mathrm{d}x \\
&= \int_{\{x \in \mathcal{X} | f_\theta(x) > 0\}} h(x) \frac{\partial}{\mathrm{d}\theta} f_\theta(x) \mathrm{d}x - \underbrace{\mathbb{E}_\theta[h(X)]}_{\text{bornée}} \underbrace{\int_{\{x \in \mathcal{X} | f_\theta(x) > 0\}} \frac{\partial}{\mathrm{d}\theta} f_\theta(x) \mathrm{d}x}_{=0}
\end{aligned}$$

$$
\begin{aligned}
&= \int_{\{x\in\mathcal{X}|f_\theta(x)>0\}} [h(x) - \mathbb{E}_\theta[h(X)]] \frac{\partial}{\mathrm{d}\theta} f_\theta(x)\mathrm{d}x \\
&= \int_{\{x\in\mathcal{X}|f_\theta(x)>0\}} [h(x) - \mathbb{E}_\theta[h(X)]] \sqrt{f_\theta(x)} \frac{\frac{\partial}{\mathrm{d}\theta} f_\theta(x)}{f_\theta(x)} \sqrt{f_\theta(x)}\mathrm{d}x \\
&\leq \sqrt{\int_{\{x\in\mathcal{X}|f_\theta(x)>0\}} [h(x) - \mathbb{E}_\theta[h(X)]]^2 f_\theta(x)\mathrm{d}x \int_{\{x\in\mathcal{X}|f_\theta(x)>0\}} \frac{\left[\frac{\partial}{\mathrm{d}\theta} f_\theta(x)\right]^2}{f_\theta^2(x)} f_\theta(x)\mathrm{d}x} \\
&\quad \text{par Cauchy-Schwartz} \\
&\leq \sqrt{\mathbb{V}_\theta[h(X)]} \sqrt{\int_{\{x\in\mathcal{X}|f_\theta(x)>0\}} \frac{\left[\frac{\partial}{\mathrm{d}\theta} f_\theta(x)\right]^2}{f_\theta(x)} \mathrm{d}x} \\
&\leq \sqrt{\mathbb{V}_\theta[\widehat{g}]} \sqrt{I_1(\theta)}.
\end{aligned}
$$

En passant au carré et en divisant par $I_1(\theta)$ (qui est strictment positif), nous avons le résultat.

Définition 44 (Borne de Cramer-Rao)**.**

Dans le cas où la quantité d'intérêt est $g(\theta)$ avec g dérivable, la **borne de Cramer-Rao** est la quantité :

$$\frac{(g'(\theta))^2}{I_1(\theta)}.$$

Corollaire 37 (Borne de Cramer-Rao et estimateur UMVU)**.**

Si un estimateur sans biais $\widehat{g}$ vérifie pour tout $\theta \in \Theta$ que sa variance est égale à la borne de Cramer-Rao alors c'est un estimateur UMVU.

Corollaire 38 (Borne de Cramer-Rao pour un n-échantillon)**.**

Étant donné un n-échantillon $\mathbf{X} = (X_1, \ldots, X_n)$ d'information de Fisher I_n et sous les conditions du théorème précédent, nous avons :

$$\forall\theta \in \Theta,\ \mathbb{V}_\theta[h(X_1, \ldots, X_n)] \geq \frac{(g'(\theta))^2}{nI_1(\theta)}.$$

Exemple

Étant donné $\mathbf{X} = (X_1, \ldots, X_n)$ un n-échantillon de loi de Poisson $\mathcal{P}(\theta^\star)$ alors nous montrons que $\widehat{\theta}_{n,1} = \overline{X}_n$ est un UMVU.

Point méthode.

Pour un estimateur $UMVU$, il faut procéder en plusieurs étapes :

1. Vérifier s'il est non biaisé (sinon, cela s'arrête là).
2. Calculer la dérivée première ou la dérivée seconde en θ de la fonction densité f_θ (suivant si nous pouvons dériver qu'une seule ou deux fois sous le signe somme).
3. Calculer l'information de Fisher pour une observation.
4. Calculer la variance de l'estimateur.
5. Vérifier si borne de Cramer-Rao est atteinte ou non.

Chapitre 4
Intervalle et région de confiance

"Capitaine, je suis sûr à 80% que c'est le meurtrier."
Adrien Monk, personnage éponyme de la série *Monk*.

4.1 Objectifs

Le but de ce chapitre est de comprendre qu'un risque est toujours possible et qu'il faut donc en tenir compte.

4.2 Introduction

Ce chapitre commence par un certain nombre de définitions.

Définition 45 (Région de confiance)**.**

Étant donné $\alpha \in]0,1[$, une **région de confiance de $g(\theta^\star)$ au niveau** $1-\alpha$ est un ensemble $\widehat{\mathcal{C}}$ construit mesurablement par rapport à l'observation X et tel que pour tout $\theta^\star \in \Theta$, nous ayons :

$$\mathbb{P}_{\theta^\star}\left(g(\theta^\star) \in \widehat{\mathcal{C}}\right) \geq 1-\alpha.$$

Lorsque l'inégalité devient une égalité, nous disons que le **niveau de confiance est exactement égal à** $1-\alpha$.

Nous avons immédiatement la généralisation pour un n-échantillon :

Définition 46 (Région de confiance asymptotique)**.**

Étant donné $\alpha \in]0,1[$, une **région de confiance asymptotique de $g(\theta^\star)$ au niveau** $1-\alpha$ est une suite d'ensembles $\left(\hat{\mathcal{C}}_n\right)_{n\in\mathbb{N}}$, chacun construit mesurablement par rapport au n-échantillon $(X_1,\ldots,X_n)$ et telle que pour tout $\theta^\star \in \Theta$, nous ayons :

$$\liminf_{n\to+\infty} \mathbb{P}_{\theta^\star}\left(g(\theta^\star) \in \hat{\mathcal{C}}_n\right) \geq 1-\alpha.$$

Remarque

Usuellement, nous prenons $\alpha = 5\%$ ou $\alpha = 1\%$.

Remarque

Étant donné un niveau de confiance fixé, une région de confiance est d'autant meilleure que son aire est petite.

Définition 47 (Intervalle de confiance)**.**

Si $g(\theta^\star) \in \mathbb{R}$, nous parlons plutôt d'**intervalle de confiance**.

⚠ Attention au piège

Une région de confiance est une variable aléatoire et nous calculons donc la probabilité que cette variable aléatoire englobe le paramètre d'intérêt qui lui est fixé. En particulier, il ne faut pas confondre la région de confiance qui est aléatoire et son estimée (sa réalisation).
Si prends ma région de confiance $\hat{\mathcal{C}}_n = [\widehat{A}_n, \widehat{B}_n]$ telle que :

$$\mathbb{P}_{\theta^\star}\left(\theta^\star \in [\widehat{A}_n, \widehat{B}_n]\right) = 95\%$$

et que je calcule une réalisation de mon intervalle, par exemple $[0.33, 0.75]$ alors on ne peut plus dire que $\mathbb{P}_{\theta^\star}\left(\theta^\star \in [0.33, 0.75]\right) = 95\%$ puisque tout est fixé :

- Soit $\theta^\star$ est dans l'intervalle et la probabilité vaut 1.
- Soit il ne l'est pas et la probabilité est nulle.

La notion de région de confiance est de dire que si nous répétons l'expérience un grand nombre de fois alors le paramètre d'intérêt se trouvera dans la région de confiance estimée en moyenne 95% du temps.

4.3 Premières constructions

La première façon de construire un intervalle de confiance est d'utiliser les quantiles :

Définition 48 (Quantile d'ordre p)**.**

Pour tout $\beta \in]0;1[$, nous appelons **quantile d'ordre** β d'une variable aléatoire X de fonction de répartition F_X est définie par

$$q_\beta = \inf \{q \in \mathbb{R} \,|\, F_X(q) \geq \beta\}.$$

Il est parfois noté $F^{-1}(\beta)$.

En particulier, nous avons :

Définitions 49 (Quantiles particuliers)**.**

- La **médiane** est le quantile d'ordre 50%.
- Les **quartiles** sont les quantiles d'ordre $25k\%$ avec $k \in \{1, 2, 3\}$.
- Les **déciles** sont les quantiles d'ordre $10k\%$ avec $k \in \{1, \ldots, 9\}$.
- Les **centiles** sont les quantiles d'ordre $k\%$ avec $k \in \{1, \ldots, 99\}$.

Remarque

Si la fonction est inversible (voir figure 4.1 (a)), le quantile est directement l'inversion de cette dernière. Si la fonction possède un plateau horizontal (voir figure 4.1 (b)), il faut prendre au début du plateau (la fonction quantile aura donc un saut). Si la fonction possède un saut (voir figure 4.1 (c)), il faut prendre la valeur correspondante à ce saut (la fonction quantile aura donc un plateau horizontal).

Exemple

Pour la loi $\mathcal{N}(0, 1)$, deux quantiles classiques (à connaître) sont : $q_{0,975} \approx 1,96$ et $q_{0,975} \approx 1,645$.

Exercices 4.1

Étant donné un n-échantillon $\mathbf{X} = (X_1, \ldots, X_n)$ de même loi $\mathcal{N}(\theta^\star, 1)$.

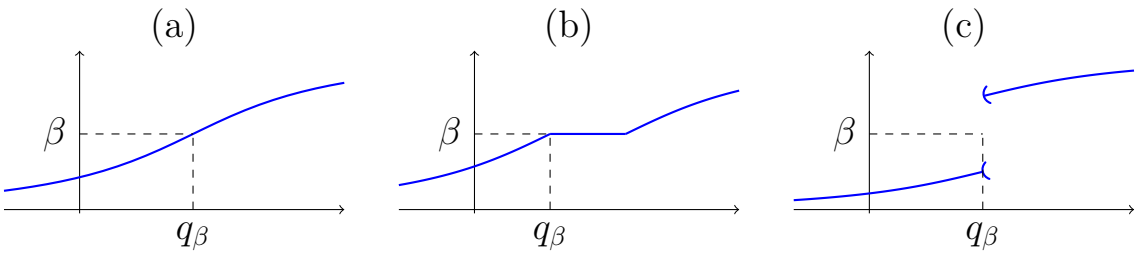


FIGURE 4.1 – Calcul de différents quantiles en fonction du type de fonction de répartition : inversible (a), avec un plateau (b) ou avec un saut (c).

Proposer un intervalle de confiance de niveau exactement 95% de la forme $]-\infty, a]$, $[a, +\infty[$ et un **intervalle bilatère** c'est-à-dire de la forme $[a, b]$. Lequel préférez-vous ?

Point méthode (Méthode du pivot)**.**

Pour calculer un intervalle de confiance, nous procédons souvent à l'aide de la **méthode dite du pivot** :

1. Choix d'un estimateur.
2. Calcul de sa loi en fonction de $\theta^\star$.
3. Transformation de l'estimateur pour obtenir une statistique dont la loi ne dépend plus de $\theta^\star$.
4. Détermination des quantiles de la loi pour estimer la région de confiance adéquat.

4.4 Intervalle de confiance de niveaux obtenus par des inégalités de probabilités

Dans cette partie, nous présentons des intervalles de confiance obtenus à partir d'inégalités de probabilités classiques.

4.4.1 Cas de variances uniformément bornée

Proposition 39 (Inégalité de Bienaymé-Tchebychev)**.**

Soit une variable aléatoire Y admettant un moment d'ordre 2 alors nous avons :

$$\forall \varepsilon > 0, \mathbb{P}\left(|Y - \mathbb{E}[Y]| \geq \varepsilon\right) \leq \frac{\mathbb{V}[Y]}{\varepsilon^2}.$$

Preuve

Ce théorème est une conséquence de l'inégalité de Markov :

Lemme 40 (Inégalité de Markov)**.**

Soit Z une variable aléatoire réelle presque sûrement positive ou nulle alors nous avons :

$$\forall a > 0, \mathbb{P}\left(Z \geq a\right) \leq \frac{\mathbb{E}[Z]}{a}.$$

Preuve

Pour tout $a > 0$, nous avons presque sûrement $Z \geq a\mathbb{1}_{\{Z \geq a\}}$ et nous avons donc :

$$\mathbb{E}[Z] \geq \mathbb{E}\left[a\mathbb{1}_{\{Z \geq a\}}\right] = a\mathbb{P}\left(Z \geq a\right) \Leftrightarrow \mathbb{P}\left(Z \geq a\right) \leq \frac{\mathbb{E}[Z]}{a}. \quad (4.1)$$

(4.2)

Comme la fonction $x \mapsto x^2$ est bijective sur $\mathbb{R}^+$ à valeurs dans $\mathbb{R}^+$, nous avons pour tout $\varepsilon > 0$:

$$\mathbb{P}\left(|Y - \mathbb{E}[Y]| \geq \varepsilon\right) = \mathbb{P}\left((Y - \mathbb{E}[Y])^2 \geq \varepsilon^2\right) \quad (4.3)$$

(4.4)

et nous pouvons appliquer l'inégalité de Markov avec $Z = (Y - \mathbb{E}[Y])^2$ et $a = \varepsilon^2$.

Exemple

Étant donné un n-échantillon de loi $\mathcal{B}(\theta^\star)$, l'estimateur du maximum

de vraisemblance est la moyenne $\overline{X}_n$ et nous avons pour tout $\varepsilon > 0$:

$$\mathbb{P}_{\theta^\star}\left(\left|\overline{X}_n - \theta^\star\right| \geq \varepsilon\right) \leq \frac{\theta^\star\left(1-\theta^\star\right)}{n\varepsilon^2} \leq \frac{1}{4n\varepsilon^2} = \alpha.$$

Donc, un intervalle de confiance bilatéral de niveau $1-\alpha$ est :

$$IC_{1-\alpha}\left(\theta^\star\right) = \left[\overline{X}_n - \frac{1}{2\sqrt{n\alpha}}, \overline{X}_n + \frac{1}{2\sqrt{n\alpha}}\right].$$

Pour $\alpha = 5\%$ et $n = 100$, la longueur est d'environ 0.45.

4.4.2 Cas de lois à supports inclus dans un compact donné

Lemme 41 (Inégalité de Hoeffding)**.**

Étant donnés, $(Y_1, \ldots, Y_n)$ des variables aléatoires réelles indépendantes et $(a_1, \ldots, a_n)$ et $(b_1, \ldots, b_n)$ des n-uplets de réels tels que pour tout $i \in \{1, \ldots, n\}$:

- $\mathbb{E}\left[Y_i\right] = 0$,
- $a_i \leq Y_i \leq b_i$ presque sûrement.

Alors, pour tout $t > 0$, nous avons :

$$\mathbb{P}\left(\sum_{i=1}^{n} Y_i > t\right) \leq 2\exp\left(-\frac{2t^2}{\sum_{i=1}^{n}\left(b_i - a_i\right)^2}\right).$$

Corollaire 42.

Étant donné un n-échantillon $\mathbf{X} = (X_1, \ldots, X_n)$ de loi commune à support inclus dans $[a, b]$ admettant un moment d'ordre 1, nous avons :

$$\forall \varepsilon > 0,\ \mathbb{P}_{\theta^\star}\left(\left|\overline{X}_n - \mathbb{E}_{\theta^\star}\left[X_1\right]\right| \geq \varepsilon\right) \leq 2\exp\left(-\frac{2nt^2}{(b-a)^2}\right).$$

Exemple

Étant donné un n-échantillon $\mathbf{X} = (X_1, \ldots, X_n)$ de loi $\mathcal{B}\left(\theta^\star\right)$ alors

nous avons presque sûrement $a = 0 \leq X_i \leq b = 1$ et donc pour tout $\varepsilon > 0$:

$$\mathbb{P}_{\theta^\star}\left(\left|\overline{X}_n - \theta^\star\right| \geq \varepsilon\right) \leq 2e^{-2nt^2}.$$

Le choix de $\varepsilon = \sqrt{1/(2n)\log(2/\alpha)}$ conduit à :

$$IC_{1-\alpha}(\theta^\star) = \left[\overline{X}_n - \sqrt{\frac{1}{2n}\log\frac{2}{\alpha}}, \overline{X}_n + \sqrt{\frac{1}{2n}\log\frac{2}{\alpha}}\right].$$

Pour $\alpha = 5\%$ et $n = 100$, la longueur est d'environ 0.27.

Remarque

De manière générale, comme la variance d'une variable aléatoire encadrée par a et b est plus petite que $(b-a)^2/4$, comparer la précision des deux méthodes revient à comparer $1/\sqrt{2\alpha}$ et $\sqrt{\log(2/\alpha)}$. Pour les petites valeurs de α, c'est l'intervalle fourni par l'inégalité de Hoeffding qui propose un meilleure résultat au prix d'une hypothèse plus forte.

4.5 Intervalle de confiance asymptotique

Le théorème de la limite centrale permet d'avoir à chaque fois un intervalle de confiance de niveau asymptotique $1-\alpha$ pour l'espérance d'une loi puisque nous avons :

$$\sqrt{n}\left(\overline{X}_n - \mathbb{E}_{\theta^\star}[X_1]\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, \mathbb{V}_{\theta^\star}[X_1]\right).$$

Nous pouvons alors proposer l'intervalle :

$$\left[\overline{X}_n \pm q_{1-\alpha/2}\sqrt{\frac{\mathbb{V}_{\theta^\star}[X_1]}{n}}\right]$$

comme intervalle de confiance (asymptotique) à condition de connaître la variance. Or, cette variance dépend quasiment tout le temps de $\theta^\star$.

Exemple

Étant donné un n-échantillon $\mathbf{X} = (X_1, \ldots, X_n)$ de loi $\mathcal{B}(\theta^\star)$ alors nous avons :

$$\sqrt{n}\left(\overline{X}_n - \theta^\star\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, \theta^\star(1-\theta^\star)\right)$$

et il faudrait isoler $\theta^\star$ mais c'est compliqué.

4.5.1 Cas de variances uniformément bornées

Dans l'exemple précédent, si nous *admettons* connaître la variance $\theta^\star(1-\theta^\star)$ alors nous pouvons proposer :

$$IC_{1-\alpha}\left(\theta^\star\right) = \left[\overline{X}_n \pm q_{1-\alpha/2}\sqrt{\frac{\theta^\star(1-\theta^\star)}{n}}\right] \subset \left[\overline{X}_n \pm \frac{q_{1-\alpha/2}}{2\sqrt{n}}\right].$$

Pour $\alpha = 5\%$ et $n = 100$, nous trouvons une longueur de 0.196.

4.5.2 Estimation consistante de la variance

S'il existe une suite d'estimateurs $\left(\widehat{\sigma}_n^2\right)_{n\in\mathbb{N}^\star}$ de la variance consistante alors d'après le lemme de Slutsky, nous avons pour tout $\theta^\star \in \Theta$:

$$\sqrt{\frac{n}{\widehat{\sigma}_n^2}}\left(\overline{X}_n - \mathbb{E}_{\theta^\star}\left[X_1\right]\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0,1\right).$$

D'où l'intervalle de confiance de niveau asymptotiquement exact $1-\alpha$:

$$IC_{1-\alpha}\left(\theta^\star\right) = \left[\overline{X}_n \pm q_{1-\alpha/2}\sqrt{\frac{\widehat{\sigma}_n^2}{n}}\right].$$

Exemple

Étant donné un n-échantillon $\mathbf{X} = (X_1, \ldots, X_n)$ de loi $\mathcal{B}\left(\theta^\star\right)$ alors nous avons :

$$IC_{1-\alpha}\left(\theta^\star\right) = \left[\overline{X}_n \pm q_{1-\alpha/2}\sqrt{\frac{\overline{X}_n\left(1-\overline{X}_n\right)}{n}}\right].$$

4.5.3 Stabilisation de la variance

À l'aide de la méthode Delta, nous avons que pour toute fonction φ continument dérivable sur $\Theta \subset \mathbb{R}$, nous avons :

$$\sqrt{n}\left(\varphi(\overline{X}_n) - \varphi\left(\mathbb{E}_{\theta^\star}\left[X_1\right]\right)\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, \varphi'^2\mathbb{V}_{\theta^\star}\left[X_1\right]\right).$$

Donc, si prenons φ telle que $\varphi'(\theta^\star) = 1/\sqrt{\mathbb{V}_{\theta^\star}\left[X_1\right]}$ alors nous aura la convergence en loi suivante :

$$\sqrt{n}\left(\varphi(\overline{X}_n) - \varphi\left(\mathbb{E}_{\theta^\star}\left[X_1\right]\right)\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0,1\right)$$

et nous pouvons en tirer l'intervalle de confiance suivant :

$$IC_{1-\alpha}\left(\theta^{\star}\right) = \left[\varphi^{-1}\left(\varphi(\overline{X}_n) \pm \frac{q_{1-\alpha/2}}{\sqrt{n}}\right)\right].$$

Exemple

Étant donné un n-échantillon $\mathbf{X} = (X_1, \ldots, X_n)$ de loi $\mathcal{B}\left(\theta^{\star}\right)$, nous devons résoudre $\varphi'(\theta^{\star}) = 1/\sqrt{\theta^{\star}(1-\theta^{\star})}$. Or, nous savons que la dérivée de la fonction arcsin est :

$$\arcsin'(x) = \frac{1}{\sqrt{1-x}}$$

alors en prenant $\varphi(x) = 2\arcsin\sqrt{x}$, nous avons :

$$IC_{1-\alpha}\left(\theta^{\star}\right) = \left[\sin^2\left(\arcsin\sqrt{\overline{X}_n} \pm \frac{q_{1-\alpha/2}}{2\sqrt{n}}\right)\right].$$

Chapitre 5

Test

"Dans la justice, il y a deux visions :

- *En France, nous sommes présumé innocent jusqu'à ce que notre culpabilité soit établie ;*
- *Aux États-Unis, nous sommes présumé coupable jusqu'à ce que notre innocence soit établie."*

5.1 Objectifs

La notion de test est assez complexe car elle repose sur une asymétrie des deux hypothèses mises en concurrence.

5.2 Formalisme et démarche expérimentale

Dans le cadre du modèle paramétrique $(E, \mathcal{E}, \mathbb{P}_\theta, \theta \in \Theta)$, nous nous donnons deux sous-ensembles Θ_0 et Θ_1 **disjoints** et inclus dans Θ (nous n'imposons pas qu'ils forment une partition). À la vue d'une observation, nous souhaitons décider si $\theta^\star \in \Theta_0$ ou pas ; si ce n'est pas le cas, nous considérerons que $\theta^\star \in \Theta_1$.

Définitions 50 (Hypothèses)**.**

Étant donnés deux sous-ensembles Θ_0 et Θ_1 **disjoints** et inclus dans Θ, nous définissons respectivement :

$\mathcal{H}_0$: l'**hypothèse nulle** considère que $\theta^\star \in \Theta_0$.

$\mathcal{H}_1$: l'**hypothèse alternative** considère que $\theta^\star \in \Theta_1$.

Pour tout $j \in \{0, 1\}$, nous dirons que $\mathcal{H}_j$ est **simple** si Θ_j est réduit à un singleton et **composite** sinon.

Exemple

Dans le cas d'une pièce, nous nous intéressons souvent à savoir si la probabilité d'avoir *pile* est égale à celle d'avoir *face* ($\Theta_0 = \{1/2\}$) ou pas ($\Theta_1 = [0,1]\backslash\{1/2\}$).

À l'aide de ces notations, nous pouvons introduire la définition d'un test :

Définitions 51 (Test)**.**

Nous appelons **test de l'hypothèse $\mathcal{H}_0$ contre l'hypothèse $\mathcal{H}_1$** toute fonction $\phi(X)$ à valeur dans $\{0,1\}$ avec ϕ mesurable et pouvant dépendre de Θ_0 et Θ_1 :

- Lorsque $\phi(X) = 0$, nous disons que nous **conservons** l'hypothèse $\mathcal{H}_0$.
- Lorsque $\phi(X) = 1$, nous disons que nous **rejetons** l'hypothèse $\mathcal{H}_0$ et nous **acceptons** l'hypothèse $\mathcal{H}_1$.

Remarque

Un test peut toujours s'écrire $\phi(X) = \mathbb{1}_{\{X\in\mathcal{R}\}}$ où $\mathcal{R} \in \mathcal{E}$ ou encore $\phi(X) = \mathbb{1}_{\{h(X)\in\mathcal{R}'\}}$ avec h mesurable et $\mathcal{R}' \in \mathcal{B}(\mathbb{R})$ la tribu des boréliens de $\mathbb{R}$.

Définitions 52 (Statistique de test)**.**

Dans le cas où $\phi(X) = \mathbb{1}_{\{h(X)\in\mathcal{R}'\}}$ avec h mesurable et $\mathcal{R}' \in \mathcal{B}(\mathbb{R})$ la tribu des boréliens de $\mathbb{R}$, nous appelons $\mathcal{R}'$ la **région de rejet** et $h(X)$ la **statistique de test**.
Le complémentaire de la région de rejet est parfois appelée **région d'acceptation**.

5.2.1 Mesure de la qualité d'un test

Pour mesurer la qualité d'un test, nous nous intéressons d'une part à la probabilité d'accepter à tord et d'autre part de rejeter à tord.

Définitions 53 (Risques de première et seconde espèce)**.**

Les **risques de première** et **de seconde espèce** du test $\phi(X)$ sont définis comme les fonctions $\underline{\alpha}$ sur Θ_0 et $\underline{\beta}$ sur Θ_1 par :

$$\underline{\alpha} : \begin{array}{ccc} \Theta_0 & \to & [0,1] \\ \theta^\star & \mapsto & \mathbb{P}_{\theta^\star}\left(\phi(X)=1\right) \end{array} \quad \text{et} \quad \underline{\beta} : \begin{array}{ccc} \Theta_1 & \to & [0,1] \\ \theta^\star & \mapsto & \mathbb{P}_{\theta^\star}\left(\phi(X)=0\right) \end{array}$$

et nous pouvons introduire la **puissance d'un test** comme la fonction $1-\underline{\beta}$ et sa **taille** comme le réel $\alpha^\star$ défini par

$$\alpha^\star = \sup_{\theta^\star \in \Theta_0} \underline{\alpha}(\theta^\star) = \sup_{\theta^\star \in \Theta_0} \mathbb{P}_{\theta^\star}\left(\phi(X)=1\right).$$

Enfin, nous disons que le test est de **niveau** α si sa taille $\alpha^\star$ est inférieure ou égale à α.

Comme pour les intervalles de confiances, la notion de test s'étend également au cas asymptotique

Définition 54 (Version asymptotique)**.**

Étant donnée une suite $(\phi_n)_{n\in\mathbb{N}}$ de tests et en notant $\alpha_n^\star$ la taille de ϕ_n, nous disons que la suite de tests est de **niveau asymptotique** α si

$$\limsup_{n\to+\infty} \alpha_n^\star \leq \alpha.$$

Point méthode (Représentation sous forme de tableau)**.**

Nous pouvons résumer les informations précédentes de la façon suivante :

$\phi(X)$ \ $\theta^\star$	Θ_0	Θ_1
0	Conservation correcte	Risque seconde espèce : $\underline{\beta}$
1	Risque première espèce : $\underline{\alpha}$	Puissance

Exemple

Étant donné un n-échantillon $\mathbf{X} = (X_1, \ldots, X_n)$ iid de loi $\mathcal{N}(\theta^\star, 1)$

avec $\theta \in \mathbb{R}$, nous voulons tester :

$$\begin{cases} \mathcal{H}_0 : \ \theta^\star \geq 1, \\ \mathcal{H}_1 : \ \theta^\star < 1, \end{cases} \text{ soit } \begin{cases} \Theta_0 = [1, +\infty[, \\ \Theta_1 =]-\infty, 1[. \end{cases}$$

L'intuition nous conduit à choisir $\phi(X) = \mathbb{1}_{\{\overline{X}_n < 1\}}$: le test est fondé sur la statistique $\overline{X}_n$ et la région de rejet $]-\infty, 1[$.

On voit que pour tout $\theta \geq 1$, $\underline{\alpha}(\theta^\star) = \mathbb{P}_{\theta^\star}\left(\overline{X}_n < 1\right) = F_{\mathcal{N}(0,1)}\left(\sqrt{n}\,(1-\theta^\star)\right)$ car $\overline{X}_n \sim \mathcal{N}\left(\theta^\star, 1/n\right)$ et pour tout $\theta < 1$, $\underline{\beta}(\theta^\star) = 1 - F_{\mathcal{N}(0,1)}\left(\sqrt{n}\,(1-\theta^\star)\right)$.

En revanche, la taille du test vaut $\alpha^\star = 1/2$ ce qui veut dire qu'on ne fait pas mieux qu'un lancer de pile ou face.

5.2.2 Dissymétrie des rôles des hypothèses $\mathcal{H}_0$ et $\mathcal{H}_1$

Si nous voulons minimiser l'erreur de première espère, il suffirait de toujours conserver l'hypothèse $\mathcal{H}_0$; à l'opposé, si nous souhaitons minimiser l'erreur de seconde espèce, il suffit de rejeter à chaque fois l'hypothèse $\mathcal{H}_0$ au profit de l'hypothèse $\mathcal{H}_1$. Nous voyons donc que minimiser les deux erreurs en même temps est compliquée (voir impossible sauf dans certains cas). Il a donc été choisi de dissymétriser le problème en faisant jouer un rôle particulier à $\mathcal{H}_0$ et en contrôlant exclusivement le risque de première espère. On prendra par exemple pour $\mathcal{H}_0$:

- Une hypothèse communément admise : par exemple, il n'y a pas de traces de vie autre que sur Terre dans notre système solaire.
- Une hypothèse de prudence ou conservatrice : le nouveau médicament qui est en train d'être testé ne soigne pas mieux que l'ancien.
- Ou plus simplement la seule hypothèse que l'on peut facilement formuler.

Dans cette vision le statisticien impose le niveau α (par exemple 5%) et acceptera de rejeter à tord son hypothèse dans 5% des cas.

5.2.3 Démarche de construction et mise en œuvre pratique

Exemple

Étant donné un n-échantillon $X_1, \ldots, X_n$ de loi normale $\mathcal{N}(\theta^\star, 1)$ et regardons le test suivant :

$$\begin{cases} \mathcal{H}_0 : \ \theta^\star \geq 1, \\ \mathcal{H}_1 : \ \theta^\star < 1. \end{cases}$$

La statistique est naturelle est toujours $\overline{X}_n$ et nous choisissons, à la vue de $\mathcal{H}_1$, une région de rejet de la forme $]-\infty; k_\alpha[$ c'est-à-dire une statistique de la forme $\phi(X) = \mathbb{1}_{\left\{\overline{X}_n < k_\alpha\right\}}$. La question revient donc au choix de k_α.

Pour ce faire, nous calculons la taille du test :

$$\begin{aligned}
\alpha^\star &= \sup_{\theta^\star \in \Theta_0} \mathbb{P}_{\theta^\star}\left(\overline{X}_n < k_\alpha\right) \\
&= \sup_{\theta^\star \geq 1} \mathbb{P}_{\theta^\star}\left(\sqrt{n}\left(\overline{X}_n - \theta^\star\right) < \sqrt{n}\left(k_\alpha - \theta^\star\right)\right) \\
&= \sup_{\theta^\star \geq 1} F\left[\sqrt{n}\left(k_\alpha - \theta^\star\right)\right] \text{ où } F \text{ est la fonction de répartition d'une gaussienne c} \\
&= F\left[\sqrt{n}\left(k_\alpha - 1\right)\right] \text{ car } F \text{ est croissante.}
\end{aligned}$$

Donc, pour que le niveau du test soit exactement α, il suffit que :

$$\begin{aligned}
\alpha = \alpha^\star &\Leftrightarrow \alpha = F\left[\sqrt{n}\left(k_\alpha - 1\right)\right] \\
&\Leftrightarrow q_\alpha = \sqrt{n}\left(k_\alpha - 1\right) \text{ où } q \text{ est le quantile d'une gaussienne centrée réduite} \\
&\Leftrightarrow \frac{q_\alpha}{\sqrt{n}} + 1 = k_\alpha.
\end{aligned}$$

Donc, la statistique de test est de la forme $\phi(X) = \mathbb{1}_{\left\{\overline{X}_n < \frac{q_\alpha}{\sqrt{n}} + 1\right\}}$ et a un niveau exactement de α.

5.2.4 Qualités d'un test

Nous recensons un certain nombre de propriétés éventuelles.

Définitions 55.

Un test est dit **sans biais** lorsque sa fonction puissance vérifie $1 - \underline{\beta} > \alpha$ sur Θ_1.
Une suite de tests est dit **consistante** s'ils sont tous de niveaux α et si, pour tout $\theta^\star \in \Theta_1$, nous avons :

$$\underline{\beta}_n(\theta^\star) \underset{n\to+\infty}{\longrightarrow} 0.$$

La **robustesse** exprime l'absence de sensibilité d'un test, éventuellement asymptotique, au modèle (où à la forme de l'observation). Entre d'autres termes, si la modélisation du phénomène est mauvaise, est-ce que nous obtenons les mêmes conclusions ?

Comme pour les estimateurs, nous pouvons chercher des tests ayant des propriétés plus intéressantes que les autres.

Définition 56 (Test uniformément plus puissant)**.**

Pour deux statistiques de test $\phi(X)$ et $\phi'(X)$ de niveau α, nous disons que $\phi(X)$ est **uniformément plus puissant (UPP)** que $\phi'(X)$ si leurs fonctions puissances $1 - \underline{\beta}$ et $1 - \underline{\beta}'$ vérifient $1 - \underline{\beta} \geq 1 - \underline{\beta}'$; c'est-à-dire que pour tout $\theta^\star \in \Theta_1$, nous avons :

$$\mathbb{P}_{\theta^\star}\left(\phi(X) = 1\right) \geq \mathbb{P}_{\theta^\star}\left(\phi'(X) = 1\right).$$

Nous disons que $\phi(X)$ est $UPP(\alpha)$

5.3 Un outil important : p-valeur

Le principe de la p-valeur (ou p-*value* en anglais) est expliquée sur la figure 5.1 : plutôt que de demander la valeur α visée par l'utilisateur, la plupart des logiciels renvoie la valeur estimée à partir de la moyenne ; il ne reste plus qu'à l'utilisateur de comparer cette valeur à la cible.

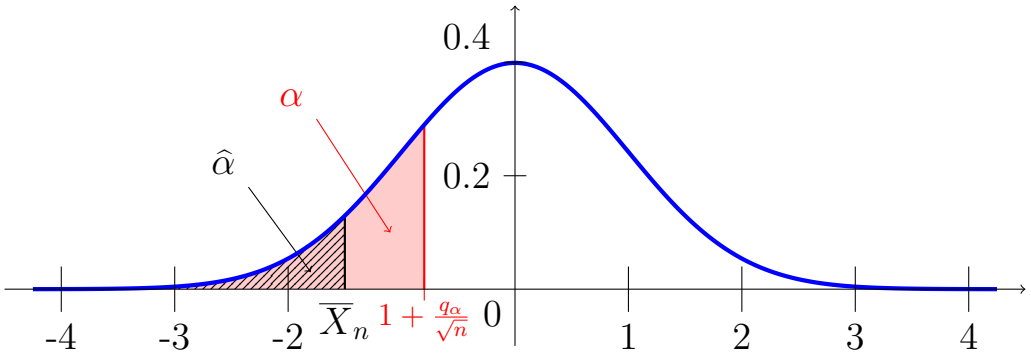


FIGURE 5.1 – Représentation schématique de la p-valeur (zone hachurée) par rapport à un niveau α visé (zone rouge).

Définition 57 (p-valeur)**.**

Supposons avoir construit une famille de test $\phi_\alpha(X)$, chacun de niveau α, pour $\alpha \in]0,1[$. La ***p*-valeur** associée à cette famille et à l'observation X est le réel défini par :

$$\widehat{\alpha}(X) = \sup\{\alpha \in]0,1[| \phi_\alpha(X) = 0\}.$$

Remarque

En d'autres termes, $\widehat{\alpha}(X)$ est le plus grand niveau autorisant l'acceptation de $\mathcal{H}_0$ et forme donc, en un certain sens, un indice de crédibilité de $\mathcal{H}_0$ à la vue de X. La p-valeur mesure l'adéquation entre $\mathcal{H}_0$ et les observations : plus $\widehat{\alpha}(X)$ est faible et plus nos informations contredisent $\mathcal{H}_0$.

Le cas typique est celui où la famille $(\phi_\alpha(X))_{\alpha\in]0,1[}$ est p.s. croissante en α, c'est-à-dire que $\phi_\alpha(X)$ décroît vers 0 lorsque α décroît vers 0 (ce qui est une condition naturelle). Ceci équivaut à dire que la famille des régions de rejet est croissante au sens de l'inclusion. Dans ce cas, on a également :

$$\widehat{\alpha}(X) = \inf\{\alpha \in]0,1[| \phi_\alpha(X) = 0\}.$$

En outre, le test ϕ_α de niveau α conserve $\mathcal{H}_0$ à la vue de X pour $\alpha < \widehat{\alpha}(X)$ et rejette $\alpha > \widehat{\alpha}(X)$.

Exemple

Étant donné un n-échantillon $\mathbf{X} = (X_1, \ldots, X_n)$ de loi $\mathcal{N}(\theta^\star, 1)$ et

si nous voulons tester :

$$\begin{cases} \mathcal{H}_0 : & \theta^\star \geq 1, \\ \mathcal{H}_1 : & \theta^\star < 1. \end{cases}$$

Nous avons vu que la région de rejet est $\overline{X}_n < k_\alpha = 1 + \frac{q_\alpha}{\sqrt{n}}$.

La p-valeur est caractérisée par le cas limite $k_{\widehat{\alpha}_n} = \overline{X}_n$ soit $\widehat{\alpha}_n(\mathbf{X}) = F\left(\sqrt{n}\left(\overline{X}_n - 1\right)\right)$.

Avec $\overline{X}_4 = 0.5$, nous obtenons $\widehat{\alpha}(\mathbf{X}) \approx 15,9\%$ et nous conservons donc l'hypothèse $\mathcal{H}_0$ avec un niveau de 10% ou même de 5%.

5.4 Botanique des tests

Dans cette partie, nous mettons quelques tests classiques.

5.4.1 Test de rapport de vraisemblance

Il s'agit d'une méthode générale pour construire des tests $UPP(\alpha)$ sous certaines hypothèses.

Définitions 58 (Statistique du rapport de vraisemblance)**.**

Étant donnés deux ensembles Θ_0 et Θ_1 représentant les deux hypothèses, nous introduisons le **rapport de vraisemblance** noté h par :

$$h(X) = \frac{\sup_{\theta\in\Theta_1} V_X(\theta)}{\sup_{\theta\in\Theta_0} V_X(\theta)}$$

où nous rappelons que $V_X(\theta) = f_\theta(X)$ est la vraisemblance du modèle. La zone de rejet naturelle est donc de la forme $]k_\alpha, +\infty[$ où k_α est à fixer suivant le contexte et la **statistique du rapport de vraisemblance** s'écrit donc $\phi(X) = \mathbb{1}_{\{h(X)>k_\alpha\}}$.

Remarque

Si la valeur $\sup_{\theta\in\Theta_0} V_X(\theta)$ est nulle, nous supposons que la statistique vaut 1 puisque le modèle 0 est rejeté. Toutefois, si le numérateur **et** le dénominateur sont nuls, il faudra remettre en cause la pertinence des hypothèses.

Exemple

Étant donné un n-échantillon $\mathbf{X} = (X_1, \ldots, X_n)$ iid de loi $\mathcal{N}(\theta^\star, 1)$ avec $\theta^\star \in \mathbb{R}$, nous voulons tester :

$$\begin{cases} \mathcal{H}_0 : \ \theta^\star \geq 1, \\ \mathcal{H}_1 : \ \theta^\star < 1. \end{cases}$$

La vraisemblance s'écrit alors :

$$V_X(\theta) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left[-\frac{1}{2}\sum_{i=1}^{n}(X_i - \theta)^2\right]$$

est donc strictement croissante sur $]-\infty, \overline{X}_n]$ et strictement décroissante sur $[\overline{X}_n, +\infty[$. Ainsi, nous avons le rapport de vraisemblance suivant :

$$h(\mathbf{X}) = \begin{cases} \frac{V_{\mathbf{X}}(\overline{X}_n)}{V_{\mathbf{X}}(1)} & \text{si } \overline{X}_n < 1, \\ \frac{V_{\mathbf{X}}(1)}{V_{\mathbf{X}}(\overline{X}_n)} & \text{si } \overline{X}_n \geq 1. \end{cases}$$

Or, pour tout θ_0 et θ_1 de $\mathbb{R}$, nous avons :

$$\frac{V_{\mathbf{X}}(\theta_1)}{V_{\mathbf{X}}(\theta_0)} = \exp\left(\frac{n}{2}\left[\theta_0^2 - \theta_1^2 + 2\overline{X}_n(\theta_1 - \theta_0)\right]\right)$$

de sorte que le rapport de vraisemblance s'écrit $h(X) = \exp\left[\frac{n}{2}\left(\overline{X}_n - 1\right)^2\left(2\mathbb{1}_{\{\overline{X}_n<1\}}\right.\right.$
Or, il s'agit d'une fonction strictement décroissante en $\overline{X}_n$ donc la statistique du rapport de vraisemblance peut se réécrire de la forme $\mathbb{1}_{\{\overline{X}_n<k_\alpha\}}$.

Théorème 43 (Lemme de Neyman-Pearson)**.**

Étant donnés un modèle paramétré par Θ et deux points θ_0 et θ_1 de Θ distincts, , nous voulons tester :

$$\begin{cases} \mathcal{H}_0 : & \theta^\star = \theta_0, \\ \mathcal{H}_1 : & \theta^\star = \theta_1. \end{cases}$$

Étant donné $\alpha \in]0,1[$, si nous supposons qu'il existe un seul k_α tel que le test de rapport de vraisemblance $\phi(X) = \mathbb{1}_{\{h(X;\theta_0,\theta_1)>k_\alpha\}}$ avec

$$h(X;\theta_0,\theta_1) = \frac{\sup\limits_{\theta\in\Theta_1} V_X(\theta)}{\sup\limits_{\theta\in\Theta_0} V_X(\theta)} = \frac{f_{\theta_1}(X)}{f_{\theta_0}(X)}$$

vérifie $\mathbb{P}_{\theta_0}(\phi(X) = 1) = \alpha$ alors ce test est $UPP(\alpha)$.

Preuve

Nous considérons qu'il existe un autre test Ψ de niveau αtel que $\mathbb{E}_{\theta_0}[\Psi(X)] = \mathbb{P}_{\theta_0}(\Psi(X) = 1) \leq \alpha$. Nous avons donc par construction $\mathbb{E}_{\theta_0}[\phi(X) - \Psi(X)] \geq 0$. Nous voulons donc montrer que le test de rapport de vraisemblance est au moins aussi puissant c'est-à-dire que

$$\mathbb{E}_{\theta_1}[\phi(X) - \Psi(X)] = \mathbb{P}_{\theta_1}(\phi(X) = 1) - \mathbb{P}_{\theta_1}(\Psi(X) = 1) \geq 0.$$

Pour ce faire, nous allons faire un changement de densité en remarquant que partout où la densité θ_0 est non nulle, nous avons

$$f_{\theta_1}(x) = \frac{f_{\theta_1}(x)}{f_{\theta_0}(x)} f_{\theta_0}(x).$$

Nous avons donc :

$$\begin{aligned}
&\mathbb{E}_{\theta_1}[\phi(X) - \Psi(X)] - k_\alpha \mathbb{E}_{\theta_0}[\phi(X) - \Psi(X)] \\
&= \mathbb{E}_{\theta_0}\left[(\phi(X) - \Psi(X))\frac{f_{\theta_1}(X)}{f_{\theta_0}(X)}\right] + \mathbb{E}_{\theta_1}\left[(\phi(X) - \Psi(X))\mathbb{1}_{\{f_{\theta_0}(X)=0\}}\right] - k_\alpha \mathbb{E}_{\theta_0}[\phi(X) \\
&= \mathbb{E}_{\theta_0}\left[(\phi(X) - \Psi(X))\left(\frac{f_{\theta_1}(X)}{f_{\theta_0}(X)} - k_\alpha\right)\right] + \mathbb{E}_{\theta_1}\left[(\phi(X) - \Psi(X))\mathbb{1}_{\{f_{\theta_0}(X)=0\}}\right] \text{ car } k_\alpha
\end{aligned}$$

Or, $\phi(X) = \mathbb{1}_{\{h(X;\theta_0,\theta_1)>k_\alpha\}}$ donc si elle vaut 0 nous avons :

$$\mathbb{E}_{\theta_0}\left[-\underbrace{\Psi(X)}_{>0}\underbrace{\left(\frac{f_{\theta_1}(X)}{f_{\theta_0}(X)} - k_\alpha\right)}_{<0}\right] > 0,$$

et si elle vaut 1, l'espérance est également positive par construction. De plus, pour les valeurs x $f_{\theta_0}(x) = 0$, l'hypothèse nulle ne peut pas être choisie donc comme k_α est fini, la statistique vaut 1. Au final, comme $k_\alpha > 0$, nous avons donc :

$$\mathbb{E}_{\theta_1}\left[\phi(X) - \Psi(X)\right] \geq k_\alpha \mathbb{E}_{\theta_0}\left[\phi(X) - \Psi(X)\right] \geq 0.$$

Chapitre 6
Rappels de bases

"On ne peut bâtir sur du sable."
Moses Isegawa

Dans ce chapitre, nous rappelons quelques théorèmes nécessaires à une bonne pratique de la statistique.

Définition 59 (Matrice jacobienne)**.**

Soit F une fonction d'un ouvert $U \subset \mathbb{R}^n$ à valeur dans $\mathbb{R}^m$ définie par ses m fonctions composantes à valeurs réelles

$$F : \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \mapsto \begin{pmatrix} f_1(x_1, \cdots, x_n) \\ \vdots \\ f_m(x_1, \cdots, x_n) \end{pmatrix}$$

alors la **matrice jabobienne de** F, notée J_F, est définie pour tout $(x_1, \ldots, x_n) \in \mathbb{R}^n$ par :

$$J_F(x_1, \ldots, x_n) = \begin{pmatrix} \frac{\partial f_1}{\mathrm{d}x_1}(x_1, \ldots, x_n) & \cdots & \frac{\partial f_1}{\mathrm{d}x_n}(x_1, \ldots, x_n) \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\mathrm{d}x_1}(x_1, \ldots, x_n) & \cdots & \frac{\partial f_m}{\mathrm{d}x_n}(x_1, \ldots, x_n) \end{pmatrix}.$$

Théorème 44 (Changement de variables dans une intégrale multiple)**.**

Soient U et V deux ouverts de $\mathbb{R}^n$, $\Phi : U \to V$ un $\mathcal{C}^1$-difféomorphisme et f une application de V dans $\mathbb{R}$. Si U et V sont bornés et si f et $(f \circ \Phi)|\det J_\Phi|$ sont Riemann-intégrables alors :

$$\int_V f = \int_U (f \circ \Phi)|\det J_\Phi|.$$

Bibliographie

B. Hauchecorne. *Les contre-exemples en mathématiques*. Ellipses Paris, 1988.

J.-F. Le Gall. Intégration, probabilités et processus aléatoires. *Ecole Normale Supérieure de Paris*, 2006.

H. Lehning. Cryptographie et codes secrets. l'art de cacher. 26, 2006.

C. Robert. *Le choix bayésien : Principes et pratique*. Springer Science & Business Media, 2006.

B. Ycart. Histoires de mathématiques. https ://hist-math.u-ga.fr/, 2017.

B. Ycart et V. Genon-Catalot. *Vecteurs et suites aléatoires. Martingales discrètes*. Centre de Publication universitaire, 2004.

Index