Rappels de Statistique Mathématique

Cours de Mastère Spécialisé Data Science

Claude PETIT, université de Rennes

14 juin 2026

Table des matières

Avant propos

Ce polycopié de statistique mathématique couvre un programme d'une trentaine d'heures de cours dispensées en début d'année de mastère spécialisé. Le niveau requis est celui d'une licence: les outils de base du calcul des probabilités sont supposés connus (théorie de la mesure, lois usuelles, fonctions caractéristiques, etc.), ainsi que les différentes notions de convergence des suites de variables aléatoires.

Le cours a deux ambitions: rappeler de façon rapide et concise les bases de la modélisation statistique, de l'estimation et des tests qui sont enseignés en première année de master ou seconde année d'école d'ingénieurs et introduire un certain nombre de notions qui seront utiles aux différents cours du mastère. Cela permettra en outre de donner un aperçu très succint des thématiques abordées dans certains d'entre eux. L'objectif est de réviser et de maîtriser les outils présentés, l'accent sera donc clairement mis sur la résolution d'exercices.

Le polycopié est un complément du cours, il ne le remplace absolument pas. Beaucoup d'exemples donnés durant les séances n'apparaissent pas dans ce document et à l'inverse beaucoup de démonstrations ou de paragraphes supplémentaires qui s'y trouvent ne seront pas traités en cours; pour ces raisons, il est donc nécessaire de prendre le cours. On trouvera également dans ce document des leçons qui traitent de sujets connexes ou d'applications des statistiques à divers domaines.

Le contenu du polycopié est très fortement inspiré du cours de statistique mathématique de Myriam Vimond et Magali Fromont, de l'ENSAI, et les exercices de travaux dirigés ont été pillés sans vergogne dans ceux des cours correspondants à l'ENSAI et à l'ENSAE (en particulier dans le cours de statistique 1 de Nicolas Chopin). Qu'ils en soient ici remerciés. Le polycopié comprend également beaucoup d'emprunts et de références aux ouvrages d'Alain Monfort « cours de statistique mathématique » et de Philippe Tassi « méthodes statistiques ».

Les parties non traitées ou les démonstrations ne sont pas exigibles en examen, mais elles pourront intéresser ceux d'entre vous qui veulent aller plus loin.

Le site web du cours de maths contient des documents supplémentaires. La bibliothèque de l'école contient également un nombre important de livres de cours et d'exercices corrigés en statistique mathématique, mais n'oubliez pas que la totalité du cours sera traitée en un mois et que vous n'aurez pas le temps de lire beaucoup d'ouvrages durant cette courte période.

Chapitre 1

Modèle statistique et échantillonnage

1.1 Modèle statistique

Une expérience aléatoire se modélise par un espace probabilisé. C'est la donnée d'un triplet $(\mathbb{X}, \mathfrak{F}, \mathbb{P})$ où $\mathbb{X}$ est l'ensemble des issues possibles de l'expérience aléatoire, $\mathfrak{F}$ est une tribu sur $\mathbb{X}$ (la tribu des évènements associés à l'expérience aléatoire) et $\mathbb{P}$ est une mesure de probabilité sur $\mathfrak{F}$.

Définition 1

Un modèle statistique $(\mathbb{X}, \mathfrak{F}, \mathscr{P})$ est la donnée de:

- L'espace $\mathbb{X}$ des observations.
- Une tribu $\mathfrak{F}$ sur $\mathbb{X}$ représentant l'ensemble des évènements sur $\mathbb{X}$.
- Une famille $\mathscr{P}$ de mesures de probabilités sur $\mathfrak{F}$.

Un modèle statistique est une famille d'expériences aléatoires ayant même espace d'observations et même tribu d'évènements. Une observation x est une réalisation d'une variable aléatoire X qui suit l'une des lois de probabilité (inconnue) de la famille $\mathscr{P}$. L'un des objectifs de l'inférence statistique est de déterminer cette loi de probabilité à partir des observations x. Vous devez donc maintenant savoir distinguer $(\mathbb{X}, \mathfrak{F}, \mathbb{P})$ et $(\mathbb{X}, \mathfrak{F}, \mathscr{P})$.

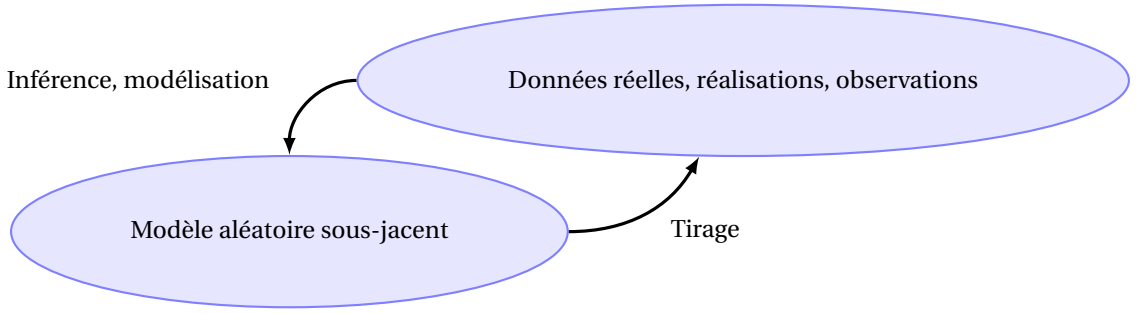

Figure 1.1 – Modélisation statistique.

Le monde réel est déterministe, notre modèle mathématique aléatoire. Il s'agit de trouver une loi de probabilité $\mathbb{P} \in \mathscr{P}$ qui modélise le mieux le phénomène observé. On notera $X \sim \mathbb{P}$ pour indiquer que X suit la loi $\mathbb{P} \in \mathscr{P}$. On rappelle qu'une variable aléatoire est une variable dont la valeur dépend du hasard. Plus rigoureusement, c'est une application mesurable $X : \mathbb{X} \longrightarrow \mathbb{R}$. Mesurable signifie que pour tout borélien B de $\mathbb{R}$, l'image réciproque $X^{-1}(B) \in \mathfrak{F}$. La mesurabilité est stable par toutes les opérations usuelles, par passage au sup et à la limite; dans notre cadre, la très grande majorité des fonctions sont mesurables.

Exemple: épreuve de Bernoulli.

Soit X une va de loi de Bernoulli $b(\theta)$, où θ est inconnu. Soit $x \in \{0,1\}$ le résultat d'un lancer à pile ou face; x est une réalisation de X. Un choix raisonnable de la famille $\mathscr{P}$ est l'ensemble des lois de Bernoulli de paramètre θ. L'espace des observations est $\mathbb{X} = \{0,1\} = \{\pi, \phi\} = \{\text{succès}, \text{échec}\}$ et $\mathfrak{F}$ est, par exemple, l'ensemble des parties de $\mathbb{X}$: $\mathfrak{F} = \{\emptyset, \{0\}, \{1\}, \{0,1\}\}$

h!

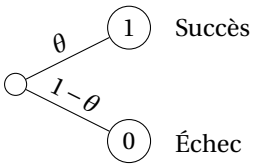

Figure 1.2 – Modèle fondamental du tirage de Bernoulli

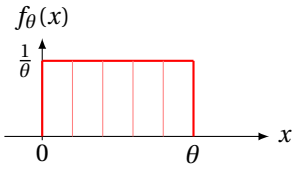

Figure 1.3 – Densité d'une loi uniforme

Exemple: loi uniforme.

Soit X une va de loi uniforme sur $[0,\theta]$, où θ est inconnu. X représente le résultat d'un tirage aléatoire équitable entre 0 et θ et tout nombre a la même probabilité d'être tiré. Soit x une réalisation de X; si $x = 0,1$, on peut en déduire que $\theta \geq 0,1$. Si $x = 2$, on peut en déduire que $\theta \geq 2$. Un choix raisonnable de la famille $\mathscr{P}$ est l'ensemble des lois uniforme sur $[0,\theta]$ pour θ variant dans $]0,\infty[$. L'espace des observations est $\mathbb{X} = [0,\theta]$ et $\mathfrak{F}$ est, par exemple, l'ensemble des boréliens de $\mathbb{X}$. $\mathbb{P}$ est défini par sa densité:

$$f(x) = \frac{1}{\theta}\mathbb{1}_{[0,\theta]}(x) \tag{1}$$

$\mathbb{P}_\theta$ est donc caractérisé par le paramètre réel θ et toute valeur de θ définit une mesure de probabilité. On notera cette famille de probabilités:

$$\mathscr{P} = \{\mathbb{P}_\theta \ ; \ \theta > 0\} \tag{2}$$

Observons n réalisations $x_1,...,x_n$ d'une v.a. X, de manière indépendante. Soient $X_1,...,X_n$ les v.a.i.i.d. sous-jacentes à ces observations. Les X_i sont à valeurs dans $(\mathbb{X},\mathfrak{F})$. Soit $\mathbb{P}$ la loi commune des X_i. $(X_1,...,X_n)$ est appelé n-échantillon de X.

Définition 2

On appelle modèle d'échantillonnage associé au n-échantillon, le triplet

$$\left(\mathbb{X}^n, \mathfrak{F}^{\otimes n}, \left\{\mathbb{P}^{\otimes n}, \mathbb{P} \in \mathscr{P}\right\}\right) \tag{3}$$

où $\mathscr{P}$ est une famille de probabilité définie sur $(\mathbb{X},\mathfrak{F})$.

Par indépendance et identique distribution des X_i, $\mathbb{P}^{\otimes n}$ est la loi du vecteur $(X_1,...,X_n)$ et l'on rappelle que la loi produit est définie par:

$$\mathbb{P}^{\otimes n}(dx_1,...,dx_n) = \prod_{i=1}^{n} \mathbb{P}(dx_i) \tag{4}$$

Exemple: répétition de n épreuves de Bernoulli indépendantes.

Le modèle d'échantillonnage associé est $\left(\{0,1\}^n, \mathfrak{F}^{\otimes n}, \left\{b(\theta)^{\otimes n}, \theta \in [0,1]\right\}\right)$, avec $\mathfrak{F} = \{\emptyset,\{0\},\{1\},\Omega\}$

Ce modèle traduit simplement le fait que dans une expérience répétée de façon indépendante, la loi de probabilité du n-échantillon est la loi produit et l'espace sous-jacent est le produit cartésien de l'espace associé à l'un des éléments de l'échantillon.

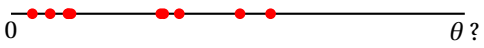

Figure 1.4 – maximum d'un échantillon uniforme : comment estimer θ ?

Supposons que $n = 10$ et mettons qu'une réalisation du n-échantillon soit $(0,0,1,1,1,0,1,0,0,1)$. Nous souhaitons évaluer θ à partir de cet échantillon. Comment peut-on procéder ?

Exemple: répétition de n épreuves d'une loi uniforme sur $[0,\theta]$.

Supposons que $n = 10$ et que nous disposions de l'échantillon suivant: $(0,1;0,6;0,2;1,2;3;0,5;1,5;2;1;1,9)$. Comment estimer θ ?

1.2 Statistique sur un modèle

Définition 3

Une statistique S sur un modèle $(\mathbb{X},\mathfrak{F},\mathscr{P})$ et à valeurs dans $(\mathbb{Y},\mathscr{G})$ est une application mesurable

$$S : \mathbb{X} \longrightarrow \mathbb{Y} \tag{5}$$
$$x \mapsto y = S(x) \tag{6}$$

Pour une réalisation x de X, $y = S(x)$ est la réalisation de la va $S(X)$.

Une statistique est donc simplement une fonction (mesurable). C'est ni plus ni moins qu'une variable aléatoire.

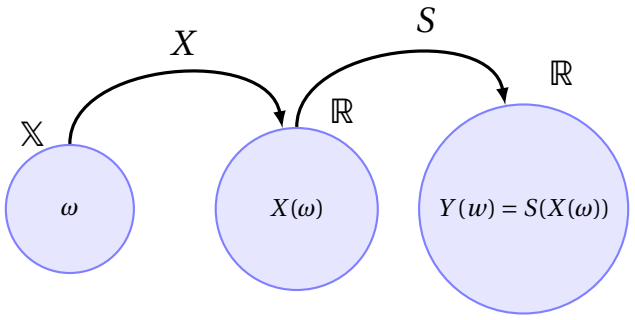

Figure 1.5 – La statistique $S : \mathbb{R} \longrightarrow \mathbb{R}$ est une fonction mesurable qui transforme X en $Y = S(X)$.

Exemple: variable binomiale et répétition de n épreuves de Bernoulli indépendantes.

$$S(x_1,x_2,...,x_n) = \sum_{i=1}^{n} x_i \tag{7}$$

Définit une statistique qui est la somme de toutes les observations du n-échantillon.

Toute statistique étant une fonction mesurable des données, elle permet de définir une variable aléatoire à partir des données initiales. La loi image $\mathbb{P}_S$ de $\mathbb{P}$ par S permet alors de définir un modèle image à partir du modèle statistique initial.

Dans l'exemple précédent, la variables aléatoire est $S(X) = \sum_{i=1}^{n} X_i$ et la loi image est la loi binomiale $\mathbb{P}_S \sim B(n,\theta)$. Le modèle image est donné par $(\mathbb{Y},\mathscr{G},\{\mathbb{P}_S : \mathbb{P} \in \mathscr{P}\})$. Considérer S plutôt que X pour mener l'inférence consiste à travailler directement dans le modèle image par S:

$$\mathbb{P}[S \in B] = \mathbb{P}[S(X) \in B] = \mathbb{P}_S(B) = \mathbb{P}[X \in S^{-1}(B)], \quad \forall B \in B_{\mathbb{R}} \tag{8}$$

Exemple: Statistique définie par le maximum d'une loi uniforme sur $[0,\theta]$.

$$S(X_1, X_2, ..., X_n) = \sup_{i=1}^{n} X_i \tag{9}$$

L'information concernant la loi inconnue de X au travers de la statistique $S(X)$ est contenue dans la tribu $\mathfrak{F}(S) \subset \mathfrak{F}(X)$, et l'on a égalité si, et seulement si, S est bijective. Habituellement, $\mathfrak{F}(S)$ est plus petite que $\mathfrak{F}(X)$ car on attend d'une statistique qu'elle réduise la tribu et condense l'information.

1.3 Quelques types de modèles statistiques

1.3.1 Rappel: mesure dominée et théorème de Radon-Nikodym

Étant donné deux mesures σ-finies μ et ν sur l'espace $(\mathbb{X}, \mathfrak{F})$, on dit que ν est absolument continue par rapport μ et l'on note $\nu \ll \mu$, si

$$\forall B \in \mathfrak{F}, \mu(B) = 0 \Rightarrow \nu(B) = 0 \tag{10}$$

On rappelle qu'en ce cas, le théorème de Radon-Nikodym assure que ν possède une densité f par rapport à μ définie par

$$\nu(B) = \int_B f(x) d\mu(x) \tag{11}$$

On note $f = d\nu/d\mu$. f est positive, mesurable et s'appelle densité ou dérivée au sens de Radon-Nikodyn de ν par rapport à μ. Par exemple, la mesure de Lebesgue est définie par $\lambda(dx) = dx$.

1.3.2 Modèle dominé

Définition 4

Un modèle statistique $(\mathbb{X}, \mathfrak{F}, \mathscr{P})$ est dominé si, et seulement si, tout $\mathbb{P} \in \mathscr{P}$ est dominé par une même mesure μ. μ est la mesure dominante du modèle.

Si un modèle est dominé par une mesure σ-finie μ, alors le modèle image est dominé par la mesure μ_S. En effet, pour tout ensemble mesurable B,

$$\mu_S(B) = 0 \iff \mu(S^{-1}(B)) = 0 \Rightarrow \left(\mathbb{P}(S^{-1}(B)) = 0 \iff \mathbb{P}_S(B) = 0\right) \tag{12}$$

1.3.3 Modèle paramétrique

Définition 5

Un modèle statistique $(\mathbb{X}, \mathfrak{F}, \mathscr{P})$ est paramétré si les éléments de $\mathscr{P}$ peuvent être décrit par un paramètre. Plus précisément, le modèle est paramétré s'il existe une surjection Λ

$$\Lambda : \Theta \longrightarrow \mathscr{P} \tag{13}$$
$$\theta \mapsto \mathbb{P}_\theta \tag{14}$$

On peut alors écrire $\mathscr{P} = \{\mathbb{P}_\theta, \theta \in \Theta\}$. Si Λ est une bijection, on dit que le modèle est identifiable.

Le modèle est paramétrique si $\Theta \subset \mathbb{R}^p$ pour un entier p donné. Sinon, il est non-paramétrique.

Exemple : épreuve de Bernoulli.

X suit de loi de Bernoulli $b(\theta)$ paramétrée par $\theta \in [0,1]$. Le modèle est donc paramétrique, identifiable car à chaque valeur de θ correspond une unique loi de Bernoulli. Ce modèle est également dominé par la mesure de comptage discrète sur $\mathbb{N}$.

Si, au lieu de considérer le paramètre θ, on choisit comme paramètre $\cos^2\theta$, le modèle n'est plus identifiable car $\mathbb{P}_\theta = \mathbb{P}_{\theta+\pi}$.

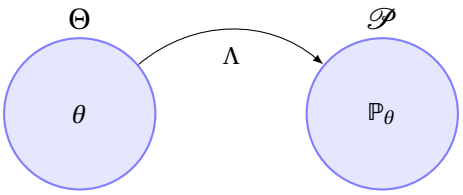

Figure 1.6 – Modèle paramétrique.

Exemple : modèle d'échantillonnage gaussien.

On considère un n-échantillon de loi gaussienne de moyenne m et écart type σ. Le modèle statistique correspondant

$$\left(\mathbb{R}^n, B(\mathbb{R})^{\otimes n}, \left\{\mathcal{N}(m,\sigma^2)^{\otimes n} : m \in \mathbb{R}, \sigma^2 > 0\right\}\right) \tag{15}$$

est paramétrique, dominé et identifiable. Il est paramétré par le couple (m,σ^2), dominé par la mesure de Lebesgue sur $\mathbb{R}^n$ et identifiable car toute loi gaussienne est caractérisée par sa moyenne et sa variance.

Exemple: modèles non paramétriques.

L'ensemble des lois de probabilités sur $\mathbb{R}$ n'est pas paramétrique. Il en est de même de l'ensemble des lois qui ont une densité par rapport à la mesure de Lebesgue ou bien l'ensemble des lois réelles continues dont la densité est symétrique. De façon générale, lorsque le paramètre n'appartient pas à un espace vectoriel de dimension finie, on parle de modèle non paramétrique.

Exemple: modèles semi-paramétriques.

On peut s'intéresser à des problèmes dépendant d'un paramètre $\theta \in \mathbb{R}^p$, mais également d'un autre paramètre appartenant à un espace de dimension infinie. Ce second paramètre représente souvent un terme de nuisance. On parle en ce cas de problème semi-paramétrique. Nous y reviendrons de façon plus précise dans le chapitre sur la robustesse.

1.3.4 Vraisemblance

Définition 6

On considère un modèle statistique $(\mathbb{X}, \mathfrak{F}, \{\mathbb{P}_\theta; \theta \in \Theta\})$ dominé par une mesure μ σ-finie. La vraisemblance du modèle est la fonction $\theta \mapsto L(x,\theta)$ définie par

$$L(x,\theta) = \frac{d\mathbb{P}_\theta(x)}{d\mu(x)} = \frac{\mathbb{P}_\theta(dx)}{d\mu(x)} \tag{16}$$

La vraisemblance est exactement la densité de la loi $\mathbb{P}_\theta$, vue comme fonction de θ.

Soient $(\mathbb{X}, \mathfrak{B}) = (\mathbb{R}, B(\mathbb{R}))$ et $P_\theta \ll \lambda$ mesure de Lebesgue sur $\mathbb{R}$; supposons que P_θ possède une densité continue $f_\theta(x)$ par rapport à la mesure de Lebesgue et plaçons-nous dans un modèle d'échantillonnage. C'est un modèle paramétrique dominé par la mesure de Lebesgue sur $\mathbb{R}^n$ et la vraisemblance de $P_\theta^{\otimes n}$ relativement à $\lambda^{\otimes n}$ est

$$L(x_1,...,x_n,\theta) = \prod_{i=1}^n f_\theta(x_i). \tag{17}$$

En effet, $d\mathbb{P}_\theta(x) = \mathbb{P}_\theta(dx) = f_\theta(x)dx = L(x,\theta)dx$.

Exemple : échantillon uniforme sur $[0,\theta]$.

$$L(x,\theta) = L(x_1,...,x_n,\theta) = \prod_{i=1}^n \left(\frac{1}{\theta}\mathbb{1}_{[0,\theta]}(x_i)\right) = \frac{1}{\theta^n}\prod_{i=1}^n \mathbb{1}_{[0,\theta]}(x_i) = \frac{1}{\theta^n}\mathbb{1}_{[0,\theta]^n}(x) \tag{18}$$

Où $x = (x_1,...,x_n)$ et la dernière indicatrice vaut 1 si, et seulement si, toutes les coordonnées x_i du vecteur x appartiennent à $[0,\theta]$.

Plaçons-nous maintenant dans le cas où $(\mathbb{X}, \mathfrak{B}) = (\mathbb{N}, \mathscr{P}(\mathbb{N}))$, où $P_\theta \ll \delta$ mesure de comptage sur $\mathbb{N}$; on a alors, pour tout sous ensemble B de $\mathbb{N}$

$$P_\theta[X \in B] = \sum_{x \in B} P_\theta[X = x] = \sum_{x \in \mathbb{N}} P_\theta[X = x]\delta_x(B) \tag{19}$$

avec $\delta_x(B) = 1$ si, et seulement si $x \in B$ et $\delta_x(B) = 0$ sinon.

Le modèle d'échantillonnage correspondant est dominé par la mesure discrète sur $\mathbb{N}^n$ et la vraisemblance de $P_\theta^{\otimes n}$ par rapport à cette mesure est donnée par

$$L(x_1, ..., x_n, \theta) = \prod_{i=1}^{n} P_\theta[X_i = x_i] \tag{20}$$

Exemple : échantillon de Bernoulli de paramètre θ.

$$L(x, \theta) = L(x_1, ..., x_n, \theta) = \prod_{i=1}^{n} \left(\theta^{x_i}(1-\theta)^{1-x_i}\mathbb{1}_{\{0,1\}}(x_i)\right) = \theta^s(1-\theta)^{n-s}\mathbb{1}_{\{0,1\}^n}(x) \tag{21}$$

où $x = (x_1, ..., x_n)$, $s = \sum_{i=1}^n x_i$ et la dernière indicatrice vaut 1 si $x_i = 0$ ou 1 pour tout i.

Attention ! Bien distinguer la densité d'une v.a. (fonction de x qui dépend de θ) de la vraisemblance d'un échantillon ou d'une v.a. (fonction de θ qui dépend de x).

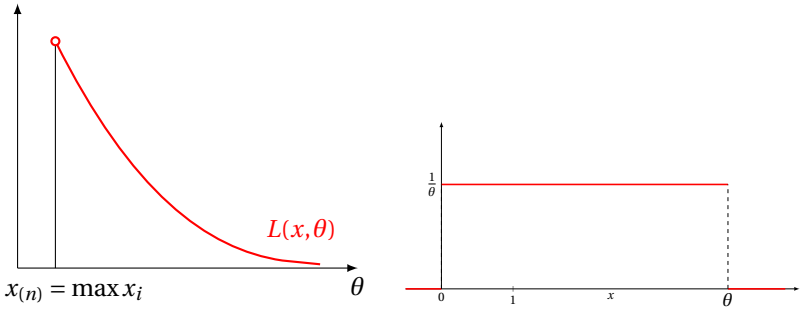

Figure 1.7 – Vraisemblance d'un échantillon uniforme (gauche) et densité d'une loi uniforme (droite).

1.3.5 Modèle homogène

Le support d'une mesure de probabilité $\mathbb{P}$ dont la densité par rapport à une mesure μ est f est par définition

$$\text{supp}(f) = \{x \in \mathbb{R} : f(x) > 0\} \tag{22}$$

Ce support n'est défini qu'à un ensemble de mesure nulle près.

Définition 7

Un modèle paramétrique est homogène si le support des lois $\mathbb{P}_\theta$ est le même.

Cela signifie également que le support ne dépend pas de θ, ou encore que toutes les lois de $\mathscr{P}$ sont équivalentes, c.-à-d. :

$$\forall \theta, \theta', \ \mathbb{P}_\theta \ll \mathbb{P}_{\theta'} \tag{23}$$

Homogène implique dominé et dans un modèle homogène, les ensembles négligeables sont les mêmes pour toutes les mesures $\mathbb{P}_\theta$. Enfin, un modèle est homogène si, et seulement si,

$$\exists\, \nu\ \sigma - \text{finie} / \ \forall \theta, \ \mathbb{P}_\theta \ll \nu \text{ et } d\mathbb{P}_\theta / d\nu > 0\ \nu - \text{p.p.} \tag{24}$$

Pour la démonstration, voir l'ouvrage d'Alain Monfort « statistique mathématique » page 36.

1.3.6 Modèle exponentiel

Définition 8

Un modèle statistique $(\mathbb{X}, \mathfrak{F}, \{\mathbb{P}_\theta; \theta \in \Theta\})$ est exponentiel si, et seulement si,

- Il est paramétrique, dominé et homogène.
- Sa vraisemblance s'écrit

$$L(x,\theta) = h(x)\exp\left(\langle\lambda(\theta), T(x)\rangle - \beta(\theta)\right) = h(x)e^{\lambda\bullet T-\beta} \tag{25}$$

h est mesurable et positive (à valeurs dans $\mathbb{R}$, $T = (T_1, ..., T_k)$ est une fonction mesurable de $\mathbb{X}$ à valeurs dans $\mathbb{R}^k$ et $\lambda(\theta) = (\lambda_1(\theta), ..., \lambda_k(\theta))$ une fonction de $\Theta \subset \mathbb{R}^p$ dans $\mathbb{R}^k$, appelée paramètre du modèle. T est la statistique privilégiée (ou naturelle) du modèle. $h(x)$ est appelé mesure de base. Enfin, β est une fonction de $\mathbb{R}^p$ à valeur dans $\mathbb{R}$, appelée fonction de log-partition. Le produit scalaire de T par λ s'écrit également

$$\lambda \bullet T = \langle\lambda(\theta), T(x)\rangle = \sum_{i=1}^{k} \lambda_i(\theta) T_i(x) \tag{26}$$

Lorsque $\lambda(\theta) = \theta$, on dit que le paramètre est canonique. En ce cas, la statistique T est dite statistique canonique du modèle. Si le modèle est identifiable, c'est à dire si $\theta \longrightarrow \lambda(\theta)$ est injective, alors on peut reparamétrer le modèle avec le nouveau paramètre $\lambda \in \Lambda = \{\lambda(\theta);\ \theta \in \Theta\}$. Lorsque la fonction $\lambda(\theta)$ n'est pas injective, on dit que la famille est surparamétrée en θ.

Une famille exponentielle est de plein rang si les $T_1(x), ..., T_k(x)$ ne sont pas linéairement dépendants (presque partout).

Exemple: épreuve de Bernoulli.

Le modèle est dominé, paramétré par θ, homogène (son support est $\{0,1\}$) et la vraisemblance s'écrit

$$L(x,\theta) = \theta^x(1-\theta)^{1-x} = \exp\left(x\ln\theta + (1-x)\ln(1-\theta)\right) \tag{27}$$

$$= \exp\left(x\ln\left(\frac{\theta}{1-\theta}\right) + \ln(1-\theta)\right) \tag{28}$$

Le modèle est donc exponentiel avec pour paramètre

$$\lambda(\theta) = \ln\left(\frac{\theta}{1-\theta}\right) \tag{29}$$

pour statistique naturelle $T(x) = x$ et $\beta(\theta) = -\ln(1-\theta)$.

Exemple : échantillon de loi uniforme sur $[0,\theta]$

La densité du n-échantillon par rapport à la mesure de Lebesgue sur $\mathbb{R}^n$ est

$$f_\theta(x_1, ..., x_n) = \frac{1}{\theta^n}\mathbb{1}_{[0,\theta]^n}(x_1, ..., x_n) \tag{30}$$

où l'indicatrice vaut 1 si, et seulement si, $0 \le x_i \le \theta$ pour tout $i = 1, ..., n$.

Le modèle n'est pas homogène, il n'est donc pas exponentiel.

Exemple : loi multinomiale

Soit $X = (X_1, ..., X_p)$ une variable de loi multinomiale de paramètres (n, θ) où n fixe, θ inconnu et

$$\Theta = \left\{\theta = (\theta_1, ..., \theta_p) \in [0,1]^p / \sum_{i=1}^{p} \theta_i = 1\right\} \tag{31}$$

Cette loi modélise le tirage de n objets appartenant à p catégories différentes, θ_i étant la proportion d'objets de catégorie i. La composante X_i du vecteur X représente alors le nombre d'objets de catégorie i parmi les n objets tirés.

Le modèle statistique modélisant un tirage de X est un modèle exponentiel. Il est dominé par la mesure de comptage sur $\mathbb{N}^p$, homogène car le support de la loi est

$$\text{supp}(X) = B_n = \left\{ x \in \mathbb{N}^p : \sum_{i=1}^{n} x_i = n \right\} = \left\{ x \in \mathbb{N}^p : s = n \right\} \tag{32}$$

et la densité (discrète) est donnée par

$$h_\theta(x) = \mathbb{P}_\theta \left[X_1 = x_1, ..., X_p = x_p \right] = \frac{n!}{x_1!...x_p!} \mathbb{1}_{B_n}(x) \prod_{i=1}^{n} \theta_i^{x_i} = \frac{n!}{x_1!...x_p!} \mathbb{1}_{B_n}(x) \exp\left(\sum_{i=1}^{p} x_i \ln \theta_i \right) \tag{33}$$

On pourrait donc choisir comme paramètre $(\ln\theta_1, ..., \ln\theta_p)$ et pour statistique $X_1, ..., X_p$, mais il s'agirait alors d'un surparamétrage car la condition $\sum X_i = n$ indique que la famille $(X_1, ..., X_p)$ est liée. Elle est en fait de rang $p-1$ et l'on préfèrera donc choisir comme statistique $T(X) = (X_1, ..., X_{p-1})$ et comme paramètre $\lambda(\theta) = \left(\ln(\theta_1/\theta_p), ..., \ln(\theta_{p-1}/\theta_p)\right)$ et $\beta(\theta) = -n\ln\theta_p$. Dans cet exemple, on a donc $k = p-1$. La représentation est dite minimale lorsque les T_i forment une famille libre. En ce cas, le cardinal de la famille $(T_i)_i$ s'appelle l'ordre de la famille.

Exemple : échantillon gaussien

On considère un n-échantillon gaussien $X_1, ..., X_n$ de loi $\mathcal{N}(m, \sigma^2)$ et l'on pose $\theta = (m, \sigma^2)$. Ce modèle est exponentiel (à montrer en exercice en déterminant $\lambda(\theta)$, $\beta(\theta)$ et $T(x)$). Idem si σ^2 est connu et que l'on pose $\theta = m$, ou bien si m et connu et que l'on pose $\theta = \sigma^2$ (en exercice).

Exemples : modèles binomial, de Poisson, gamma

Un modèle binomial de paramètres (n, θ) est exponentiel et la statistique X est le paramètre naturel (dans le modèle d'échantillonnage correspondant, $S_n = \sum X_i$ est le paramètre naturel). Il est en de même dans un modèle de Poisson de paramètre θ.

Considérons un modèle de loi gamma de paramètres (α, β), dont la densité est:

$$\frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \mathbb{1}_{\mathbb{R}_+}(x) \tag{34}$$

alors le paramètre naturel est donné par la statistique $(X, \ln X)$ (le prouver).

Propriétés des familles exponentielles.

Comme $L(x, \theta)$ est une densité de probabilité, un calcul rapide montre que

$$\beta(\theta) = \ln \int_{\mathbb{X}} h(x) e^{\lambda \bullet T} d\mu(x) \tag{35}$$

Cette formule justifie le nom de fonction log-partition (par analogie avec la physique statistique) et fait apparaître une formule proche d'une transformée de Laplace. Nous pouvons définir une nouvelle mesure μ^* à partir de μ en posant

$$\frac{d\mu^*}{d\mu}(x) = h(x) \tag{36}$$

de sorte que $\mu^* \ll \mu$ et la vraisemblance associée est

$$L^*(x, \theta) = \exp\left(\langle \lambda(\theta), T(x) \rangle - \beta(\theta) \right) \tag{37}$$

$L^*(x, \theta)$ est également une densité de probabilité.

Nous supposons maintenant que (quitte à reparamétrer) θ est le paramètre canonique du modèle. Les développements précédents nous amènent à poser les définitions suivantes:

Définition 9

Étant donné une mesure μ σ-finie sur $\mathbb{R}^k$ (donc en particulier une mesure de probabilité), on définit sa transformée de Laplace par

$$L_\mu(\theta) = \int_{\mathbb{X}} e^{\langle \theta, x \rangle} d\mu(x) \tag{38}$$

Le sous-ensemble de $\mathbb{X}$ sur lequel cette transformée de Laplace existe (l'intégrale doit converger) est un cône convexe et s'appelle l'espace canonique des paramètres. Il est non vide si, et seulement si, μ est une mesure de Radon (finie sur tout compact). Si μ n'est pas concentrée sur un hyperplan, la fonction $\theta \longrightarrow \ln L_\mu(\theta)$

est strictement convexe et est une fonction analytique. Si deux mesures ont des transformées de Laplace qui coïncident sur une partie de l'intérieur de leur cône convexe, alors elles sont égales (toutes ces propriétés sont admises).

Notons $\Theta(\mu)$ l'intérieur (au sens topologique) de l'espace canonique des paramètres. Alors

$$d\mathbb{P}_{\theta,\mu}(x) = \exp\left(\langle\theta, x\rangle - \ln L_\mu(\theta)\right) d\mu(x) \tag{39}$$

définit une famille de probabilité sur $\mathbb{X}$. C'est la famille engendrée par μ et de paramètre canonique (ou naturel) θ. Cette famille est homogène et dominée (par μ). Les familles engendrées par deux mesures μ_1 et μ_2 coïncident s'il existe $a \in \mathbb{R}^n$ (en supposant que $\mathbb{X} = \mathbb{R}^n$) et $b \in \mathbb{R}$ tels que

$$d\mu_1(x) = e^{<a,x>+b} d\mu_2(x) \tag{40}$$

de sorte que l'on peut engendrer la famille par n'importe lequel de ses éléments.

Le modèle image par la v.a. T est un modèle exponentiel admettant pour densité $e^{<\theta,t>-\beta(\theta)}$ (il suffit d'écrire la définition pour s'en convaincre). Si l'on considère maintenant la fonction génératrice ϕ des moments de la v.a. $T(X)$, on a par définition

$$\phi(s) = \mathbb{E}_\theta\left[e^{<s,T(X)>}\right] = \int h(x) e^{<\theta+s,T(x)>-\beta(\theta)} d\mu(x) = e^{\beta(\theta+s)-\beta(\theta)} \tag{41}$$

où $s \in \mathbb{R}^k$. Les propriétés des fonctions génératrices (la dérivation en θ sous le signe somme est licite) donnent alors le théorème suivant, qui permet le calcul des moments de $T(X)$ à l'aide des dérivées de la fonction de log-partition (à condition d'avoir un paramètre canonique $\lambda(\theta) = \theta$).

Théorème 10

L'espace canonique des paramètres est convexe et pour tout θ à l'intérieur de cet espace, les moments de la statistique canonique T peuvent s'exprimer par

$$\mathbb{E}\left[\prod_{i=1}^{k} T_i^{n_i}\right] = e^{-\beta(\theta)} \frac{\partial^{n_1+\ldots+n_k}}{\partial\theta_1^{n_1}\ldots\partial\theta_k^{n_k}} e^{\beta(\theta)} \tag{42}$$

En particulier,

$$\mathbb{E}[T_i] = \frac{\partial}{\partial\theta_i}\beta(\theta) \tag{43}$$

$$\mathbb{V}(T_i) = \frac{\partial^2}{\partial\theta_i^2}\beta(\theta) \tag{44}$$

$$\mathrm{cov}\left(T_i, T_j\right) = \frac{\partial^2}{\partial\theta_i\partial\theta_j}\beta(\theta) \tag{45}$$

Si le paramètre n'est pas sous forme canonique, on peut tout de même calculer les moments de la statistique T, sous certaines conditions: supposons $\Theta \subset \mathbb{R}$ ouvert et $\lambda(\theta)$ bijective. Supposons également que λ et β soient deux fois dérivables sur Θ. Alors:

$$\mathbb{E}_\theta[T(X)] = \frac{\beta'(\theta)}{\lambda'(\theta)} \tag{46}$$

$$\mathbb{V}_\theta(T(X)) = \frac{1}{\lambda'(\theta)^3}\left[\beta''(\theta)\lambda'(\theta) - \beta'(\theta)\lambda''(\theta)\right] \tag{47}$$

En effet, la bijectivité de la fonction λ permet d'exprimer ce paramètre en fonction de β en posant

$$\lambda = \beta(\lambda^{-1}(\theta)) \tag{48}$$

Puis pour s assez petit pour que $s + \lambda$ appartienne à Θ, de définir la fonction génératrice des moments dont la dérivée en 0 donne les résultats ci-dessus.

Exemple: loi de Bernoulli

Nous avons vu que $\lambda(\theta) = \ln\left(\frac{\theta}{1-\theta}\right)$, soit encore $\theta = e^\lambda/(1+e^\lambda)$. La fonction de log-partition, après reparamétrage, devient

$$\widetilde{\beta}(\lambda) = \lambda(\theta) = \ln(1+e^\lambda) \tag{49}$$

Sa dérivée première vaut $e^\lambda/(1+e^\lambda) = \theta$ qui est bien l'espérance d'une loi de Bernoulli de paramètre θ et sa dérivée seconde vaut $\theta(1-\theta)$ (à vérifier) qui est bien égale à la variance.

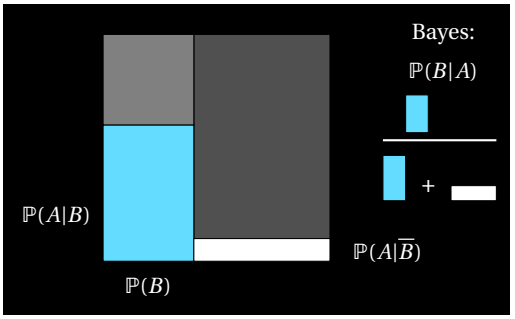

Figure 1.8 – Illustration de la formule de Bayes.

1.3.7 Modèle bayésien

Soit $\mathcal{O}$ une tribu sur l'ensemble Θ des paramètres.

Définition 11

Un modèle bayésien est un modèle statistique $(\mathbb{X}, \mathfrak{F}, \{\mathbb{P}_\theta; \theta \in \Theta\})$ muni d'une loi de probabilité Π sur l'espace $(\Theta, \mathcal{O})$

Π s'appelle la loi *a priori*. Dans un tel modèle, une valeur θ du paramètre est la réalisation d'une variable aléatoire T de loi Π.

On dispose alors de trois lois: la loi *a priori* Π, la vraisemblance $L(x|\theta)$ qui est également la loi conditionnelle de X sachant θ (loi conditionnelle des effets sachant les causes) et la loi *a posteriori* notée $L(\theta|x)$ (loi des causes sachant les effets).

Notons T une v.a. de loi Π. La formule de Bayes appliquées aux différentes densités (cette formule est admise) donne, pour presque tout x:

$$f_{T/X}(\theta, x) = \frac{f_{X/T}(x,\theta) f_T(\theta)}{f_X(x)} \tag{50}$$

Propriété

La vraisemblance *a posteriori* $L(\theta|x)$ est proportionnelle au produit $L(x|\theta) \times \Pi(\theta)$ de la vraisemblance et de la densité *a priori*.

Le coefficient de proportionnalité est la densité marginale $f_X(x)$ de X.

Exemple:

Considérons un n-échantillon $X = (X_1, ..., X_n)$ de loi de Bernoulli de paramètre θ inconnu. Cette expérience modélise le lancer de n tirages à pile ou face dont on souhaite déterminer la probabilité θ de faire pile. On pense que la pièce est truquée et des observations passées amènent à définir la loi *a priori* de θ comme une loi béta de paramètres (a, b). La vraisemblance de l'échantillon s'écrit

$$L(x, \theta) = \prod_{i=1}^{n} L(x_i, \theta) = \theta^s (1-\theta)^{n-s} \tag{51}$$

avec $s = \sum_{i=1}^{n} x_i$. D'après la formule de Bayes, la vraisemblance *a posteriori* sera proportionnelle à

$$L(\theta|x) \propto \theta^s (1-\theta)^{n-s} \frac{1}{B(a,b)} \theta^{a-1} (1-\theta)^{b-1} \mathbb{1}_{[0,1]}(\theta) \tag{52}$$

À une constante multiplicative près, il s'agit de la densité d'une loi béta de paramètres $(s+a, n-s+b)$.

Exemple:

Considérons un n-échantillon gaussien $X = (X_1, ..., X_n)$ de loi $\mathcal{N}(\theta, \sigma^2)$ dont nous souhaitons estimer le paramètre inconnu θ (la variance σ^2 est supposée connue). La loi *a priori* sur la moyenne θ est une loi gaussienne $\mathcal{N}(a, b^2)$ également, d'hyperparamètres a et b^2:

$$\Pi(\theta) = \left(2\pi b^2\right)^{-1/2} \exp\left(-\frac{1}{2b^2}(\theta - a)^2\right) \tag{53}$$

La vraisemblance de l'échantillon est

$$L(x|\theta) = \left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \theta)^2\right) \tag{54}$$

La loi jointe de (X, H) a pour densité

$$f_{(X,H)}(x,\theta) = \left(2\pi b^2\right)^{-1/2}\left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\frac{\theta^2}{2}\left(\frac{1}{b^2} + \frac{n}{\sigma^2}\right) + \theta\left(\frac{a}{b^2} + n\frac{\overline{x}}{\sigma^2}\right) - \left(\frac{a^2}{2b^2} + \frac{n\overline{x^2}}{2\sigma^2}\right)\right) \tag{55}$$

où $\overline{x} = (\sum x_i)/n$ et $\overline{x^2} = (\sum x_i^2)/n$. Posons τ et m tels que:

$$\begin{cases} \dfrac{1}{\tau^2} = \dfrac{1}{b^2} + \dfrac{n}{\sigma^2} \\ \dfrac{m}{\tau^2} = \dfrac{a}{b^2} + \dfrac{n\overline{x}}{\sigma^2} \end{cases} \tag{56}$$

Alors la loi jointe a pour forme

$$f_{(X,H)}(x,\theta) = \gamma \exp(-\Lambda/2) \tag{57}$$

où

$$\gamma = \left(2\pi b^2\right)^{-1/2}\left(2\pi\sigma^2\right)^{-n/2} \tag{58}$$

et

$$\Lambda = \frac{1}{\tau^2}\left(\theta^2 - 2m\theta\right) + \frac{a^2}{b^2} + \frac{n\overline{x^2}}{\sigma^2} \tag{59}$$

Finalement la loi jointe peut s'écrire

$$f_{(X,H)}(x,\theta) = \gamma \times \exp\left(\frac{m^2}{2\tau^2} - \frac{a}{2b^2} - \frac{n\overline{x^2}}{2\sigma^2}\right) \times \exp\left(-\frac{1}{2\tau^2}(\theta - m)^2\right) \tag{60}$$

$$= \text{constante} \times C(x) \times \mathcal{N}(m, \tau^2) \tag{61}$$

L'expression $C(x)$ ne dépend pas de θ. La densité de X va être de la forme $C(x) \times (2\pi\tau^2)^{-1/2}$ et finalement la vraisemblance *a posteriori* suit une loi $\mathcal{N}(m, \tau^2)$. m et τ^2 peuvent s'écrire de la façon suivante:

$$\tau^2 = \left(\frac{1}{b^2} + \frac{n}{\sigma^2}\right)^{-1} = \frac{b^2\sigma^2/n}{\sigma^2/n + b^2} \tag{62}$$

et

$$m = \tau^2\left(\frac{a}{b^2} + \frac{n\overline{x}}{\sigma^2}\right) = \frac{b^2\sigma^2/n}{\sigma^2/n + b^2} \times \left(\frac{a}{b^2} + \frac{n\overline{x}}{\sigma^2}\right) = \frac{\sigma^2/n}{\sigma^2/n + b^2}a + \frac{b^2}{\sigma^2/n}\overline{x} = (1-\lambda)a + \lambda\overline{x} \tag{63}$$

$$\tag{64}$$

où $\lambda \in [0,1]$ est une pondération entre la variance des observations et celle de la loi *a priori*. Les expressions de m et τ montrent qu'au fur et à mesure que le nombre d'éléments dans l'échantillon augmente, le poids des observations est de plus en plus important par rapport à l'*a priori*.

Il existe bien d'autres types de modèles statistiques: modèles dynamiques, modèles de régressions, etc. (voir, par exemple, le cours d'Alain Monfort « Statistical Methods of Econometrics » ou son ouvrage « mathématique statistique » aux éditions Economica).

Chapitre 2

Estimation ponctuelle

2.1 Définitions et propriétés des estimateurs

2.1.1 Estimateur, biais et erreur quadratique

On se place dans un modèle statistique paramétrique $(\mathbb{X}, \mathfrak{F}, \mathscr{P})$ avec $\mathscr{P} = \{\mathbb{P}_\theta, \theta \in \Theta\}$ et l'on considère une variable aléatoire $X \sim \mathbb{P}$.

Définition 13

Un estimateur S (ou S_n) de θ est une statistique à valeurs dans un sur-ensemble de Θ.

Un estimateur doit être une fonction des seules données et ne pas dépendre du paramètre θ qu'il est censé estimer. On attend bien sûr d'un estimateur qu'il donne une valeur proche de la vraie valeur de θ...

On utilisera souvent les notations S_n ou T_n (S ou T) pour un estimateur (en omettant parfois le n, mais en se souvenant que l'estimateur dépend de l'échantillon, donc de n).

Définition 14

Le biais d'un estimateur S est la quantité

$$b(\theta) = \mathbb{E}_\theta[S] - \theta \tag{1}$$

L'erreur quadratique moyenne de l'estimateur est

$$\mathbb{E}_\theta[(S-\theta)^2] = ||S-\theta||_2^2 = \mathbb{V}_\theta(S) + b(\theta)^2 \tag{2}$$

Si $b(\theta) \neq 0$, on dit que l'estimateur est biaisé.

Exemple:

Considérons un n-échantillon $X_1, ..., X_n$ d'une va X d'espérance m et de variance σ^2, sa moyenne empirique $\overline{X} = (\sum_{i=1}^n X_i)/n$ et sa variance empirique:

$$S_n^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2 \tag{3}$$

$\overline{X}$ et S_n^2 sont des estimateurs respectifs de la moyenne m et de la variance σ^2 de X.

On voit facilement que $\mathbb{E}[\overline{X}] = m$ et un calcul rapide (à faire) donne $\mathbb{E}[S_n^2] = \frac{n-1}{n}\sigma^2$.

$\overline{X}$ est donc un estimateur sans biais de m tandis que S_n^2 est un estimateur biaisé de la variance.

2.1.2 Propriétés asymptotiques et théorèmes limites

Considérons un estimateur S_n de θ dans un modèle d'échantillonnage. On dispose alors d'un échantillon $X_1, ..., X_n$ de v.a.i.i.d. de loi $\mathbb{P}_\theta$ et S_n est une fonction $S(X_1, ..., X_n)$ de $(X_1, ..., X_n)$, de sorte que S_n dépend de n. On dit que:

S_n est asymptotiquement sans biais si

$$\lim_{n\to+\infty} \mathbb{E}_\theta[S_n] = \theta \tag{4}$$

S_n converge en moyenne quadratique si

$$S_n \xrightarrow{L^2} \theta \tag{5}$$

S_n est fortement consistant pour θ si

$$S_n \xrightarrow{\mathbb{P}_\theta\text{-p.s.}} \theta \tag{6}$$

S_n est consistant pour θ si

$$S_n \xrightarrow{\mathbb{P}_\theta} \theta \tag{7}$$

On rappelle en annexe les différentes notions de convergence d'une suite de variables aléatoires, leurs différentes notations et leurs propriétés principales. En particulier, S_n converge en moyenne quadratique si, et seulement si,

$$\lim_{n\to+\infty} ||S_n - \theta||_2 = \lim_{n\to+\infty} \mathbb{E}\left[|S_n - \theta|^2\right] = 0 \tag{8}$$

De même S_n est fortement consistant pour θ si

$$\forall \theta,\ \mathbb{P}_\theta\left[\lim_{n\to+\infty} S_n = \theta\right] = 1 \tag{9}$$

De la même façon, S_n est consistant pour θ si, et seulement si,

$$\forall \epsilon > 0,\ \lim_{n\to+\infty} \mathbb{P}_\theta\left[|S_n - \theta| > \epsilon\right] = 0 \tag{10}$$

On note également $S_n = \theta + o_{\mathbb{P}}(1)$. Dire que $X_n = o_{\mathbb{P}}(1)$ signifie que X_n tend vers 0 en probabilité (sous $\mathbb{P}$).

La consistance forte implique la consistance et la convergence en moyenne quadratique implique la consistance.

On rappelle la condition pour qu'une suite de v.a. $(S_n)_n$ converge en loi vers une v.a. S. On notera $\rightsquigarrow$ ce type de convergence. Si F_n et F indiquent respectivement la fonction de répartition de S_n et de S, alors

$$S_n \rightsquigarrow S \iff \lim_{n\to+\infty} F_n(x) = F(x),\ \forall x \text{ où } F \text{ est continue.} \tag{11}$$

On rappelle sans démonstration les deux théorèmes (importants) suivants:

Théorème 15 — Loi forte des grands nombres

Soient $X_1, ..., X_n$ un n-échantillon d'une v.a. X de loi $\mathbb{P}$ où $\mathbb{P}$ admet un moment d'ordre 1 noté $\mathbb{E}[X] = m$. Alors :

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{\mathbb{P}\text{-p.s.}} m \tag{12}$$

La vitesse de convergence vers la limite est précisée par le théorème de la limite centrale.
On parle de v_n-consistance pour indiquer que v_n est la vitesse de convergence. Dire que S_n est v_n-consistant revient à dire que la suite est bornée en probabilité, c.-à-d. que :

$$v_n(S_n - \theta) = \mathcal{O}_{\mathbb{P}}(1) \tag{13}$$

Théorème 16 — de la limite centrale

Soient $X_1, ..., X_n$ un n-échantillon d'une v.a. $X \sim \mathbb{P}$ où $m = \mathbb{E}[X]$ et $\sigma^2 = \mathbb{V}(X)$ existent. Alors

$$\sqrt{n}(\overline{X} - m) \rightsquigarrow \mathcal{N}\left(0, \sigma^2\right) \tag{14}$$

De façon équivalente, on a :

$$(\overline{X} - m)/(\sigma/\sqrt{n}) \rightsquigarrow \mathcal{N}(0,1) \tag{15}$$

On dit que l'estimateur S_n est v_n-consistant et asymptotiquement normal si

$$v_n(S_n - \theta) \rightsquigarrow \mathcal{N}(0, \Sigma_\theta) \tag{16}$$

Σ_θ est une matrice $p \times p$ symétrique et positive (p est la dimension de θ); c'est la matrice de covariance de la loi limite.

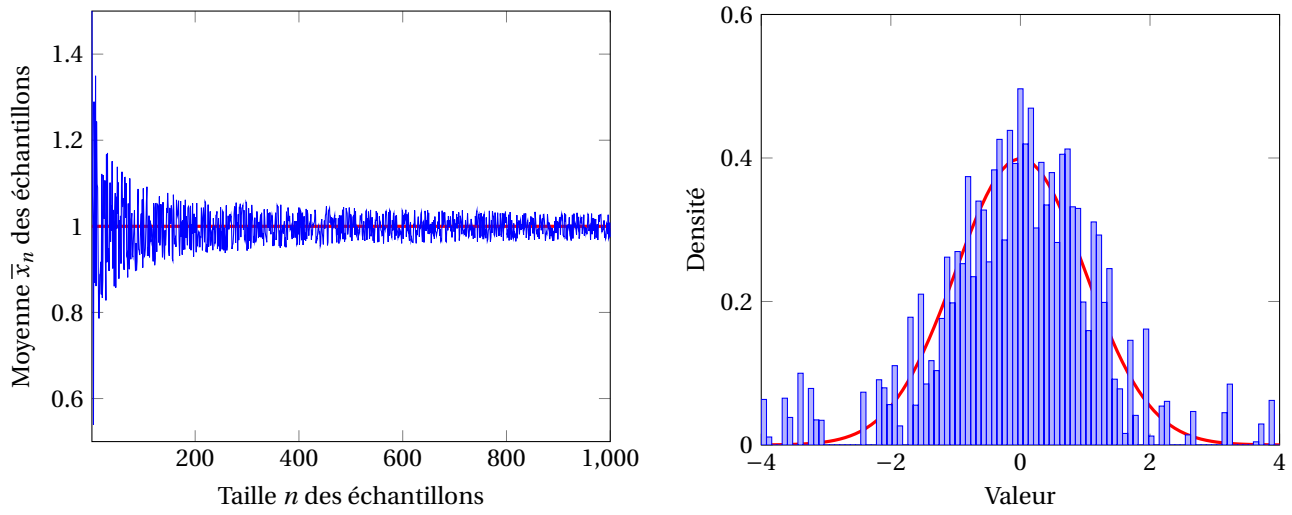

Figure 2.1 – Gauche : illustration de la LFGN: un échantillon i.i.d. de taille n est généré et sa moyenne empirique $\overline{x}_n$ est calculée pour des valeurs de n de plus en plus grande. Droite: illustration du TLC : pour $n = 100$, comparaison de la densité théorique d'un échantillon avec la densité gaussienne standard (en rouge).

2.2 Estimateurs plug-in

2.2.1 Fonction de répartition empirique

Nous supposons ici que $\mathbb{X} = \mathbb{R}$.

Considérons un n-échantillon de $X \sim \mathbb{P} \in \mathscr{P}$. Nous ne connaissons pas cette loi et l'une des questions importante de l'inférence statistique est de savoir si, lorsque la taille de l'échantillon tend vers l'infini, nous allons être capable de déterminer $\mathbb{P}$, en accumulant de plus en plus d'observations.

Trouver $\mathbb{P}$ est équivalent à trouver F, la fonction de répartition de la loi.

Définition 17

La mesure empirique des observations du n-échantillon est la loi de probabilité

$$\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i} \tag{17}$$

La fonction de répartition empirique du n-échantillon est la fonction de répartition de la loi empirique. Elle est définie par

$$F_n(x) = \mathbb{P}_n(]-\infty, x]) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{[X_i \leq x]} \tag{18}$$

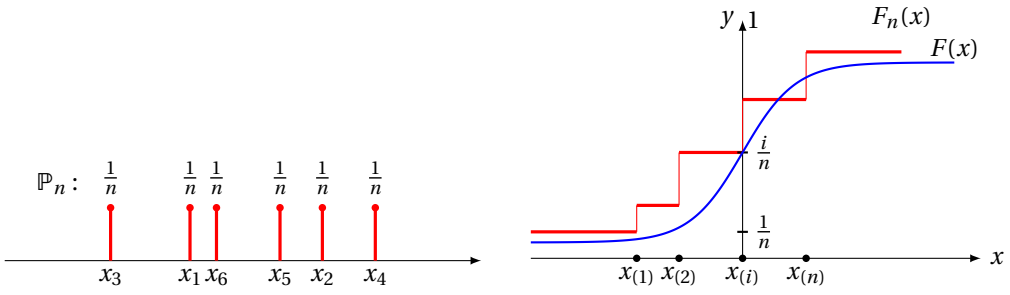

Figure 2.2 – Mesure et fonction de répartition empirique

$\mathbb{P}_n$ est une mesure aléatoire, en ce sens que pour tout borélien B, $\mathbb{P}_n(B)$ est une fonction des X_i, donc une v.a. C'est une mesure discrète qui associe un poids de $1/n$ à l'observation X_i.

F_n est une fonction croissante, continue à droite et admettant une limite à gauche (on dira « cadlag »).

Notons $X_{(i)}$ la $i$ème statistique d'ordre de l'échantillon (où $X_{(1)}$ est le minimum des $X_1, ..., X_n$ et $X_{(n)}$ le maximum). On rappelle la formule essentielle suivante:

$$\mathbb{E}[\mathbb{1}_A] = \mathbb{P}(A) \tag{19}$$

Alors:

$$F_n(x) = \begin{cases} 0 \text{ si } x \leq X_{(1)} \\ i/n \text{ si } X_{(i)} \leq x < X_{(i+1)} \\ 1 \text{ si } x \geq X_{(n)} \end{cases} \tag{20}$$

Pour tout t, n et i, on a égalité des évènements suivants:

$$\left[X_{(i)} \leq x\right] = [nF_n(x) \geq i] \tag{21}$$

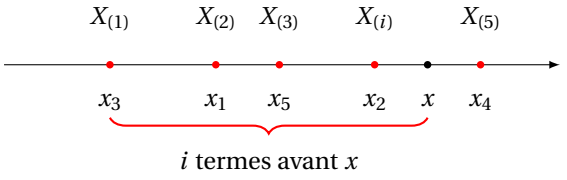

Figure 2.3 – Statistique d'ordre

Il est clair que $F_n(x)$ est un estimateur de $F(x) = \mathbb{P}[X \leq x]$. En effet, $\mathbb{E}[F_n(x)] = F(x)$ et $\mathbb{V}(F_n(x)) = F(x)(1 - F(x))/n$. $nF_n(x)$ est la somme de n v.a.i.i.d. de loi de Bernoulli de paramètre $F(x)$ et leur somme suit donc une loi binomiale de paramètres n et $F(x)$. La loi forte des grands nombres assure alors que, pour tout x fixé, $F_n(x)$ converge vers $F(x)$ presque sûrement et le théorème de la limite centrale donne le résultat suivant:

Théorème 18

$$\forall x \in \mathbb{R},\ \sqrt{n}\,(F_n(x) - F(x)) \rightsquigarrow \mathcal{N}\,(0; F(x)(1 - F(x))) \tag{22}$$

2.2.2 Le théorème fondamental de la statistique

Dans théorème le précédent, la convergence a lieu pour tout x réel *fixé*. Le théorème suivant assure que la convergence presque sûre de la fonction de répartition empirique vers la fonction de répartition de la loi limite est uniforme en x.

Théorème 19 de Glivenko-Cantelli

$$||F_n - F||_\infty = \sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \overset{\mathbb{P}_\theta - \text{p.s.}}{\longrightarrow} 0 \tag{23}$$

Ce théorème est très important (c'est le théorème fondamental de la statistique) car il assure qu'en prenant de plus en plus d'observations, la mesure empirique converge bien vers la loi limite de la vraie variable aléatoire.

Démonstration:

Soit Y une v.a. réelle de fonction de répartition G et soit U une v.a. de loi uniforme sur $[0,1]$. Soit

$$\Lambda_n = ||F_n - F||_\infty = \sup_{x \in \mathbb{R}} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{[X_i \leq x]} - F(x) \right| = \sup_{x \in \mathbb{R}} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{[F^{-1}(U_i) \leq x]} - F(x) \right| \tag{24}$$

car X_i et $F^{-1}(U_i)$ ont même loi pour tout i.

$$\Lambda_n = \sup_{x \in \mathbb{R}} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{[U_i \leq F(x)]} - F(x) \right| = \sup_{s \in F(\mathbb{R})} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{[U_i \leq s]} - s \right| \tag{25}$$

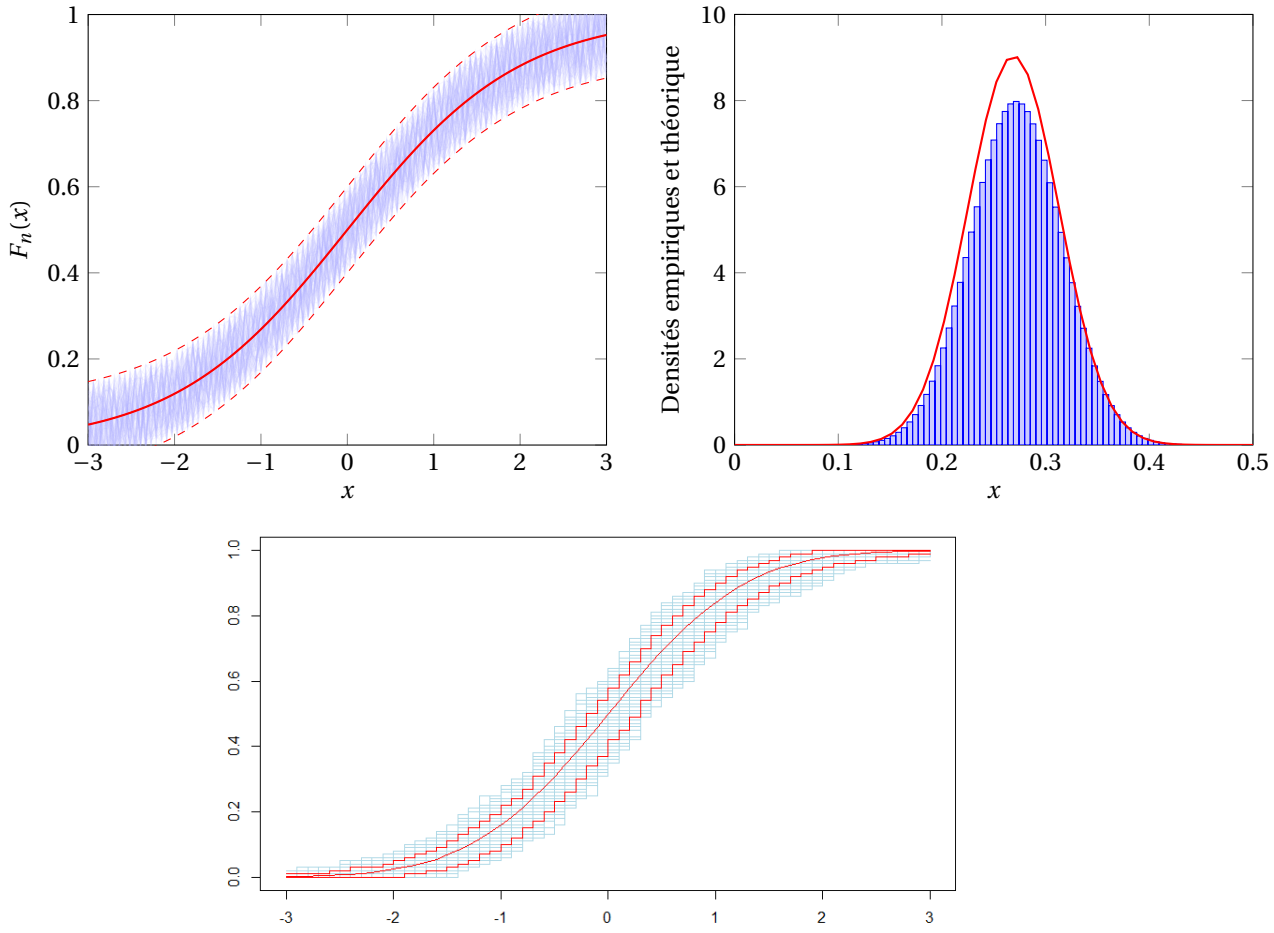

Figure 2.4 – haut, gauche: fonctions F_n (bleu) et F (rouge) avec bande de confiance uniforme. En haut à droite: illustration de la normalité asymptotique. Bas: détail de la convergence de quelques F_n vers F (crédit Arthur Charpentier).

Les U_i sont des v.a.i.i.d. de L^1 et la loi forte des grands nombres s'applique donc (pour chaque s fixé) : il existe $N_s \subset \Omega$ tel que $\mathbb{P}(N_s) = 0$ et $\forall \omega \notin N_s$,

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{[U_i(\omega)\leq s]} \overset{n\to+\infty}{\longrightarrow} s \tag{26}$$

Soit $N = \bigcup_{s\in[0,1]\cap\mathbb{Q}} N_s$. On a $\mathbb{P}(N) = 0$ et $\forall \omega \notin N,\ \forall s \in [0,1]\cap\mathbb{Q}$, la convergence précédente est vérifiée. La fonction

$$\phi_\omega(s) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{[U_i(\omega)\leq s]} \tag{27}$$

est clairement positive et croissante. Elle converge donc simplement vers s quand n tend vers l'infini, et ce $\forall \omega \notin N$. Le lemme de Dini assure alors que la convergence est uniforme (le lemme de Dini est rappelé en annexe). □

La vitesse de convergence dans le théorème de Glivenko-Cantelli est précisée par le:

Théorème 20 de Kolmogoroff-Smirnov

$$\sqrt{n}||F_n - F||_\infty \rightsquigarrow K \tag{28}$$

où K est une variable aléatoire de fonction de répartition

$$\mathbb{P}[K \leq x] = 1 - 2\sum_{k=1}^{+\infty} (-1)^{k+1} e^{-2k^2x^2} \tag{29}$$

Démonstration:

La preuve origniale est due à Kolmogorov. Une autre preuve a été donnée par Doob. En voici simplement l'idée directrice: pour tout $x \in [0,1]$, on pose $D_n(x) = \sqrt{n}(F_n(x) - F(x))$. Cette fonction aléatoire est nulle en 0 et en 1, et possède n sauts de discontinuités de hauteur $1/n$. D'après le théorème de la limite centrale (multidimensionnel), si $0 \leq x_1 \leq ... \leq x_k < 1$, la suite de vecteurs $(D_n(x_1), ..., D_n(x_k))_{n\in\mathbb{N}}$ converge en loi vers un vecteur gaussien centrée de matrice de covariance $V = (x_i(1-x_j))_{ij}$. La suite $(D_n)_n$ converge alors vers un processus continu appelé pont brownien. On peut montrer que le résultat subsiste lorsque l'on passe à la norme infinie.

□

Cette loi sert à tabuler le test de Kolmogorov-Smirnov (nous reviendrons dessus dans le chapitre sur les tests statistiques). On peut par ailleurs montrer que si F est continue, alors la loi de $||F_n - F||_\infty$ est indépendante de F (c'est une statistique libre). On dispose aussi de l'inégalité uniforme suivante (admise), appelée inégalité de Dvoretzky-Kiefer-Wolfowitz:

Théorème 21 — Inégalité DKW

$$\forall \epsilon > 0,\ \mathbb{P}\left[||F_n - F||_\infty > \epsilon\right] \leq 2e^{-2n\epsilon^2} \tag{30}$$

2.2.3 Notion de fonctionnelle statistique

Définition 22

Une fonctionnelle statistique ϕ est une fonction de $\mathbb{P}$. Elle est linéaire si elle est de la forme

$$\phi(\mathbb{P}) = \int a(x)d\mathbb{P}(x) = \mathbb{E}_{\mathbb{P}}[f(X)] \tag{31}$$

La forme la plus générale de fonctionnelle est $h\left(\mathbb{E}_{\mathbb{P}}[f(X)]\right)$ pour h et f fonctions données.

$$\phi : \mathscr{P} \longrightarrow \mathbb{R} \tag{32}$$

$$\mathbb{P} \mapsto \phi(\mathbb{P}) \tag{33}$$

ϕ est une forme « linéaire » sur un ensemble de probabilités (attention, $\mathscr{P}$ n'est pas un espace vectoriel). On parle de fonctionnelle de Von Mises (1932)

Exemples:

La moyenne: $m(\mathbb{P}) = \int x\mathbb{P}(dx)$

La variance: $\sigma^2(\mathbb{P}) = \int (x - m(\mathbb{P}))^2\mathbb{P}(dx)$

La médiane: $m_e(\mathbb{P}) = F^{-1}(1/2)$ avec $F^{-1}(p) = \inf\{t \in \mathbb{R} : F(t) > p\}$ inverse généralisée de F.

Définition 23

Un estimateur « plug in » est obtenu par substituant $\mathbb{P}_n$ à $\mathbb{P}$ dans une fonctionnelle statistique: $T = \phi(\mathbb{P}_n)$.

Pour une statistique linéaire $\phi(\mathbb{P}) = \int f(x)\mathbb{P}(dx)$, l'estimateur plug-in associé est

$$\phi(\mathbb{P}_n) = \frac{1}{n}\sum_{i=1}^{n} f(X_i) \tag{34}$$

2.2.4 Propriétés des estimateurs plug in

Si $a \in L^1(\mathbb{P})$, c'est à dire si $\int |f(x)|d\mathbb{P} = \mathbb{E}[|f(X)|] < \infty$, alors la loi forte des grands nombres s'applique et

$$\phi(\mathbb{P}_n) \overset{\mathbb{P}-\text{p.s.}}{\longrightarrow} \phi(\mathbb{P}) \tag{35}$$

Si $f \in L^2(\mathbb{P})$, c'est à dire si $\int |f(x)|^2 d\mathbb{P} = \mathbb{E}[f(X)^2] < \infty$, alors le théorème de la limite centrale s'applique et

$$\sqrt{n}(\phi(\mathbb{P}_n) - \phi(\mathbb{P})) \rightsquigarrow \mathcal{N}\left(0, \tau(\mathbb{P})^2\right) \tag{36}$$

avec $\tau(\mathbb{P})^2 = \int (f(x) - \phi(\mathbb{P}))^2 d\mathbb{P}$

2.3 Estimateur par substitution

Définition 24

Soit T un estimateur de θ et $\psi : \Theta \longrightarrow \mathbb{R}^d$ une fonction mesurable.
Un estimateur par substitution de $\psi(\theta)$ est un estimateur de la forme $\psi(T)$.

Théorème 25 de l'application continue

Soit T un estimateur (fortement) consistant de θ.
Soit $\psi : \Theta \subset \mathbb{R}^p \longrightarrow \mathbb{R}^d$ une fonction continue.
Alors $\psi(T)$ est (fortement) consistant pour $\psi(\theta)$.

Démonstration:

• $T = T_n$ est un estimateur fortement consistant si, et seulement si, $T_n \overset{\mathbb{P}_\theta-\text{p.s.}}{\longrightarrow} \theta$.

Il existe alors un ensemble Ω tel que $\mathbb{P}_\theta(\Omega) = 1$ et $\forall \omega \in \Omega$, $\lim_{n\to+\infty} T_n(\omega) = \theta$. Par continuité de ψ, on a $\lim_{n\to+\infty} \psi(T_n(\omega)) = \psi(\theta)$, $\forall \omega \in \Omega$. Ainsi, $\psi(T_n) \overset{\mathbb{P}_\theta-\text{p.s.}}{\longrightarrow} \psi(\theta)$.

• Pour montrer que la consistance de T_n vers θ implique celle de $\psi(T_n)$ vers $\psi(\theta)$, nous allons montrer de façon plus générale que si T_n converge vers T en probabilité, alors $\psi(T_n)$ converge vers $\psi(T)$ en probabilité:

Soit $\epsilon > 0$. Quelque soit δ, soit B_δ l'ensemble des x tel qu'il existe y vérifiant $||x-y|| < \delta$ et $||\psi(x)-\psi(y)|| > \epsilon$. Si $T \notin B_\delta$ et $||\psi(T_n)-\psi(T)|| > \epsilon$, alors $||T_n - T|| \geq \delta$. Notons $\beta_\delta = [X \in B_\delta]$. On a alors:

$$A_\epsilon = \left[||\psi(T_n)-\psi(T)|| > \epsilon\right] = \{\omega \in \Omega : ||\psi(T_n)-\psi(T)|| > \epsilon\} \tag{37}$$

$$A_\epsilon = \left(A_\epsilon \cap \beta_\delta\right) \cup \left(A_\epsilon \cap \overline{\beta_\delta}\right) \tag{38}$$

$$\mathbb{P}(A_\epsilon) \leq \mathbb{P}\left(A_\epsilon \cap \beta_\delta\right) + \mathbb{P}\left(A_\epsilon \cap \overline{\beta_\delta}\right) \tag{39}$$

$$\mathbb{P}(A_\epsilon) \leq \mathbb{P}\left(\beta_\delta\right) + \mathbb{P}\left[||T_n - T|| \geq \delta\right] \tag{40}$$

car $[||T_n - T|| \geq \delta] \subset A_\epsilon \cap \overline{\beta_\delta}$

Le second terme tend vers 0 quand n tend vers l'infini, pour tout δ fixé. Le premier terme tend vers 0 lorsque δ tend vers 0 en décroissant, car la suite d'ensembles $(B_\delta)_\delta$ décroit vers l'ensemble vide (par continuité de ψ).

• La propriété est également vraie pour la convergence en loi; c'est alors une conséquence du théorème de Portmanteau[1]. □

Théorème 26 Méthode delta en dimension 1

Soit T_n un estimateur de θ tel que $v_n(T_n - \theta) \rightsquigarrow T$
Soit $\psi : \Theta \subset \mathbb{R} \longrightarrow \mathbb{R}$ dérivable. Alors

$$v_n\left(\psi(T_n) - \psi(\theta)\right) \rightsquigarrow \psi'(T) \tag{41}$$

Si T_n A.N. avec $v_n(T_n - \theta) \rightsquigarrow \mathcal{N}(0, \sigma^2)$, alors $\psi(T_n)$ A.N. et

$$v_n\left(\psi(T_n) - \psi(\theta)\right) \rightsquigarrow \mathcal{N}\left(0, \Psi'(\theta)^2\sigma^2\right) \tag{42}$$

1. Portmanteau veut dire « fourre tout » en anglais. Patrick Billingsley a écrit un ouvrage célèbre sur les convergences de variables aléatoires: « Convergence of Probability Measures ». Dans cet ouvrage, il attribue le théorème à Jean-Pierre Portmanteau, de l'université de Felletin, petite commune française de la creuse. L'article est daté de 1915 et intitulé « espoir pour l'ensemble vide ». Il s'agit d'un canular. Le théorème a en fait été prouvé par Alexandre Alexandrov vers 1940. Alexandrov était le directeur de thèse de Grigori Perelman, qui a obtenu la médaille Fields pour avoir démontré la conjecture de Poincaré. Il a refusé cette médaille, ainsi que le prix Wolf d'un million de dollars qui l'accompagnait.

Théorème 27 — Méthode delta multidimensionnelle

Soit T_n un estimateur de θ tel que $v_n(T_n-\theta) \rightsquigarrow T$
Soit $\psi : \Theta \subset \mathbb{R}^p \longrightarrow \mathbb{R}^d$ une fonction différentiable de matrice jacobienne J. Alors

$$v_n\left(\psi(T_n)-\psi(\theta)\right) \rightsquigarrow d\psi_\theta(T) \tag{43}$$

et si T_n est asymptotiquement normal avec

$$v_n(T_n-\theta) \rightsquigarrow \mathcal{N}(0,\Sigma_\theta) \tag{44}$$

alors $\psi(T_n)$ est asymptotiquement normal et

$$v_n\left(\psi(T_n)-\psi(\theta)\right) \rightsquigarrow \mathcal{N}\left(0, J\times\Sigma_\theta\times J^T\right) \tag{45}$$

Remarques et cas de la dimension 1:

$d\psi_\theta(h)$ est la différentielle de ψ au point θ, calculée en h. Lorsque $d=1$, la différentielle est égale à $\psi'(\theta)\times h$. Si nous notons $J_\psi(\theta)$ la matrice jacobienne de ψ en θ, alors $d\psi_\theta(h) = J_\psi(\theta)\times h$.

$J = J_{\psi(\theta)} \in \mathbb{R}^{d\times p}$ est la matrice jacobienne de ψ et $J^T = J^T_{\psi(\theta)}$ sa transposée. Dans le cas de la dimension 1, on a simplement $J\Sigma J^T = \psi'(\theta)^2\times\sigma^2_\theta$, où σ^2_θ est la variance de la loi normale asymptotique.

Exemple:

Soit $(X_1,\dots,X_n)$ un n-échantillon d'une v.a. X de loi de Bernoulli de paramètre $\theta\in]0,1[$. Soit $\overline{X}_n$ la moyenne empirique de l'échantillon. D'après le théorème de la limite centrale,

$$\sqrt{n}\left(\overline{X}_n-\theta\right) \rightsquigarrow \mathcal{N}(0,\theta(1-\theta)) \tag{46}$$

Posons $\psi(\theta) = 2\arcsin\sqrt{\theta}$. $\psi(\overline{X}_n)$ est un estimateur de $\psi(\theta)$ et puisque ψ est dérivable, alors $\psi(\overline{X}_n)$ est $\sqrt{n}$-consistant et asymptotiquement normal.

Comme

$$\psi'(\theta) = \frac{1}{\sqrt{\theta(1-\theta)}} \tag{47}$$

Alors $J\Sigma J' = \psi'(\theta)^2\times\theta(1-\theta) = 1$ et l'on a donc

$$\sqrt{n}\left(2\arcsin\sqrt{\overline{X}_n}-2\arcsin\sqrt{\theta}\right) \rightsquigarrow \mathcal{N}(0,1) \tag{48}$$

Exemple:

Soit $(X_1,\dots,X_n)$ un n-échantillon d'une v.a. X de loi exponentielle (à la française) de paramètre $\theta>0$, que l'on souhaite estimer. Soit $\overline{X}_n$ la moyenne empirique de l'échantillon et soit $\lambda = 1/\theta$. On sait que

$$\mathbb{E}[X] = \lambda \text{ et } \mathbb{V}(X) = \lambda^2 \tag{49}$$

D'après le théorème de la limite centrale, $\sqrt{n}\left(\overline{X}_n-\lambda\right) \rightsquigarrow \mathcal{N}(0,\lambda^2)$

Posons $\psi(x) = 1/x$ qui est différentiable sur $]0,+\infty[$ avec $\psi'(x)^2 = 1/x^4$. En appliquant la méthode delta, on en déduit que

$$\frac{\sqrt{n}}{\theta}\left(\frac{1}{\overline{X}_n}-\theta\right) \rightsquigarrow \mathcal{N}(0,1) \tag{50}$$

Première démonstration de la méthode Δ:

L'idée de la démonstration consiste à effectuer un développement limité de ψ à l'ordre 1 au voisinage de θ, puis en remplaçant dans le reste $R(h)$ la variable h par la variable aléatoire $T_n-\theta$ de montrer que ce reste tend vers 0 en probabilité. Nous utilisons deux lemmes démontrés en annexe:

Lemme 1: Soient $(v_n)_n$ une suite réelle positive tendant vers l'infini et $(X_n)_n$ une suite de v.a. Si $(v_nX_n)_n$ converge en loi, alors $(X_n)_n$ converge en probabilité vers 0. □

Lemme 2: Soit $(X_n)_n$ une suite de v.a. convergeant en probabilité vers 0 et ϕ une fonction vérifiant $\phi(h) = o(||h||^p)$ quand h tend vers 0 pour $p \geq 0$. Alors $\phi(X_n) = Y_n||X_n||^p$ pour une suite $(Y_n)_n$ tendant vers 0 en probabilité. □

Si ψ est différentiable, alors

$$\psi(\theta + h) - \psi(\theta) = d\psi_\theta(h) + o(||h||) \tag{51}$$

Supposons l'existence d'une suite v_n tendant vers l'infini et d'une suite d'estimateurs T_n tels que

$$v_n(T_n - \theta) \rightsquigarrow T \tag{52}$$

Soit R la fonction représentant le reste dans le développement limité à l'ordre 1 de ψ en θ.

$$R(h) = \psi(\theta + h) - \psi(\theta) - d\psi_\theta(h) \tag{53}$$

Par différentiabilité de ψ, $R(h) = o(||h||)$ et d'après le lemme 2,

$$R(T_n - \theta) = \psi(T_n) - \psi(\theta) - d\psi_\theta(T_n - \theta) = Y_n||T_n - \theta|| \tag{54}$$

Pour une suite de v.a. $(Y_n)_n$ tendant vers 0 en probabilité. En multipliant par v_n, il vient

$$v_n\left(\psi(T_n) - \psi(\theta)\right) - d\psi_\theta(v_n(T_n - \theta)) = v_n Y_n||T_n - \theta|| \tag{55}$$

D'après le lemme de Slutsky, le terme de droite tend vers 0 en loi, donc également en probabilité. De nouveau grâce au lemme de Slutsky, $v_n\left(\psi(T_n) - \psi(\theta)\right)$ et $d\psi_\theta(v_n(T_n - \theta))$ ont même limite en loi et par continuité cette limite vaut $d\psi_\theta(T)$.

□

Seconde démonstration de la méthode Δ:

Le principe est le même que dans la première démonstration, mais la démonstration est simplifiée par l'introduction de la notion de tension et l'utilisation des o stochastiques.

Si ψ est différentiable, alors

$$\psi(\theta + h) - \psi(\theta) = d\psi_\theta(h) + o(||h||) \tag{56}$$

Supposons l'existence d'une suite v_n tendant vers l'infini et d'une suite d'estimateurs T_n tels que

$$v_n(T_n - \theta) \rightsquigarrow T \tag{57}$$

Alors $T_n = \theta + o_{\mathbb{P}}(1)$ (la suite est uniformément tendue, d'après le théorème de Prohorov) et d'après le lemme de Slutsky, $T_n - \theta$ converge vers 0 en probabilité. Soit R la fonction représentant le reste dans le développement limité à l'ordre 1 de ψ en θ.

$$R(h) = \psi(\theta + h) - \psi(\theta) - d\psi_\theta(h) \tag{58}$$

Par différentiabilité de ψ, $R(h) = o(||h||)$. Les propriétés des o stochastiques (rappelées en annexe) permettent alors d'écrire que

$$R(T_n - \theta) = \psi(T_n) - \psi(\theta) - d\psi_\theta(T_n - \theta) = o_{\mathbb{P}}(||T_n - \theta||) \tag{59}$$

car $v_n(T_n - \theta)$ est tendue. En multipliant par v_n, il vient

$$v_n\left(\psi(T_n) - \psi(\theta)\right) - d\psi_\theta(v_n(T_n - \theta)) = o_{\mathbb{P}}(1) \tag{60}$$

donc, l'écart entre $v_n\left(\psi(T_n) - \psi(\theta)\right)$ et $d\psi_\theta(v_n(T_n - \theta))$ tend vers 0 en probabilité. Par ailleurs, la multiplication (matricielle) est continue, et par le théorème de l'application continue, $d\psi_\theta(v_n(T_n - \theta)) \rightsquigarrow d\psi_\theta(T)$ et $d\psi_\theta(v_n(T_n - \theta))$ tend vers 0 en probabilité. Finalement, le lemme de Slutsky montre que $v_n\left(\psi(T_n) - \psi(\theta)\right) \rightsquigarrow d\psi_\theta(T)$.

En appliquant le théorème au cas où T est gaussien, on obtient que $\sqrt{n}(\psi(T_n) - \psi(\theta)) \rightsquigarrow \mathcal{N}(0, J \times \Sigma_\theta \times J^T)$.

□

Premier exemple en dimension 2:

On suppose que $T_n = (X_n, Y_n)^T$ vérifie $\sqrt{n}(T_n - m) \rightsquigarrow \mathcal{N}(0, \Sigma)$ avec $m = (0,1)^T$ et $\Sigma = Id$. Quelle est la loi asymptotique de $X_n Y_n$?

On pose $\Psi(x,y) = xy$. Cette fonction est de classe C^1 sur $\mathbb{R}^2$ et sa matrice jacobienne est $J = \psi\left(\frac{\partial g}{\partial x}, \frac{\partial g}{\partial y}\right) = (y,x)$. Appliquée en m, on a $J(m) = (1,0)$ et par suite

$$J(m) \times \Sigma \times J(m)^T = 1 \tag{61}$$

On en déduit donc que $\sqrt{n}X_nY_n$ tend en distribution vers une variable de loi $\mathcal{N}(0,1)$.

Second exemple en dimension 2:

Considérons un n-échantillon $(X_1,...,X_n)$ d'une v.a. X telle que $\mathbb{E}[X^4] < \infty$. On note $\alpha_i = \mathbb{E}[|X|^i]$ pour $i = 1,..,4$ et $\alpha = (\alpha_1,\alpha_2)'$. On note également

$$\overline{X} = \tfrac{1}{n}\sum_{i=1}^n X_i \quad , \quad \overline{X^2} = \tfrac{1}{n}\sum_{i=1}^n X_i^2 \quad \text{et} \quad S^2 = \tfrac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2 \tag{62}$$

Soit $T_n = \left(\overline{X}, \overline{X^2}\right)^T$. Soit $\psi(x,y) = y - x^2$. Alors $S^2 = \psi(\overline{X}, \overline{X^2})$. Le théorème de la limite centrale multidimensionnel s'applique et donne

$$\sqrt{n}(T_n - \alpha) = \sqrt{n}\left(\begin{pmatrix} \overline{X} \\ \overline{X^2} \end{pmatrix} - \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix}\right) \rightsquigarrow \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \alpha_2 - \alpha_1^2 & \alpha_3 - \alpha_1\alpha_2 \\ \alpha_3 - \alpha_1\alpha_2 & \alpha_4 - \alpha_2^2 \end{pmatrix}\right) \tag{63}$$

La fonction ψ est polynomiale donc différentiable et sa différentielle en (x,y) est $d\psi_{(x,y)}(h,k) = -2xh + k$. La méthode delta appliquée à ψ donne alors

$$\sqrt{n}\left(S^2 - (\alpha_2 - \alpha_1^2)\right) \rightsquigarrow \mathcal{N}\left(0, -4\alpha_1^4 - \alpha_2^2 + 8\alpha_1^2\alpha_2 - 4\alpha_1\alpha_3 + \alpha_4\right) \tag{64}$$

Cette loi normale est la loi de la variable aléatoire $-2\alpha_1 T_1 + T_2$ où $(T_1,T_2)'$ est le vecteur limite de $\sqrt{n}(T_n - \alpha)$.

Lorsque les variables sont centrées, c'est à dire lorsque $\alpha_1 = 0$, le résultat devient

$$\sqrt{n}\left(S^2 - \alpha_2\right) \rightsquigarrow \mathcal{N}\left(0, \alpha_4 - \alpha_2^2\right) \tag{65}$$

Exemple:

Soit $(X_1,Y_1),...,(X_n,Y_n)$ un n-échantillon d'un vecteur gaussien (X,Y) dont le coefficient de corrélation linéaire est noté ρ. Le coefficient de corrélation empirique est par définition égal à

$$\widehat{\rho}_n = \frac{\sum_{i=1}^n \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\left[\sum_{i=1}^n \left(X_i - \overline{X}\right)^2 \times \sum_{i=1}^n \left(Y_i - \overline{Y}\right)^2\right]^{1/2}} \tag{66}$$

Le théorème de la limite centrale et la méthode delta montrent que $\sqrt{n}(\widehat{\rho}_n - \rho)$ est asymptotiquement normale. On peut montrer (à faire en exercice) que

$$\sqrt{n}(\widehat{\rho}_n - \rho) \rightsquigarrow \mathcal{N}\left(0, (1-\rho^2)^2\right) \tag{67}$$

La méthode de stabilisation de la variance amène à chercher une fonction différentiable ψ dont la dérivée vaut $(1-\rho^2)^{-1}$. On pose donc naturellement $\psi(\rho) = \operatorname{arctanh}\rho$ et une nouvelle application de la méthode delta montre que

$$\sqrt{n}(\psi(\widehat{\rho}_n) - \psi(\rho)) \rightsquigarrow \mathcal{N}(0,1) \tag{68}$$

2.4 Estimateur des moments

Étant donné une variabe aléatoire X, on note $m_k(\theta) = \mathbb{E}_\theta\left[X^k\right]$ son moment théorique d'ordre k et

$$\widehat{m}_k(\theta) = \frac{1}{n}\sum_{i=1}^n X_i^k \tag{69}$$

le moment empirique d'ordre k d'un n-échantillon de X.

Définition 28

L'estimateur des moments de $\theta \in \Theta \subset \mathbb{R}^p$ est la solution $\hat{\theta}$ du système à p équations, d'inconnue θ:

$$\begin{cases} m_1(\theta) = \widehat{m}_1(\theta) \\ \vdots \\ m_p(\theta) = \widehat{m}_p(\theta) \end{cases} \quad \text{ou encore, de façn matricielle } M(\theta) = \widehat{M}(\theta) \tag{70}$$

θ est un vecteur à p coordonnées; dans le cas où $p = 1$, l'estimateur des moments est obtenu en identifiant moyenne théorique et empirique.

Exemple:

Considérons un n-échantillon $(X_1, ..., X_n)$ de X suivant une loi de Poisson de paramètre θ inconnu. On sait que $\mathbb{E}[X] = \theta$. On va donc poser comme estimateur des moments de θ la statistique $\overline{X} = \widehat{m}_1(\theta)$.

Exemple:
Considérons un n-échantillon $(X_1, ..., X_n)$ de X suivant une loi exponentielle (à la française) de paramètre θ inconnu. On sait que $\mathbb{E}[X] = 1/\theta$. On va donc poser comme estimateur des moments de θ la statistique T_n telle que

$$\overline{X} = \widehat{m}_1(\theta) = \frac{1}{T_n} \Rightarrow T_n = \frac{1}{\overline{X}} \tag{71}$$

Exemple:
L'estimateur des moments d'un n-échantillon de loi uniforme sur $[0, \theta]$ où θ est le paramètre inconnu est donné par la statistique $T_n = 2\overline{X}$ (à faire en exerice).

Exemple:
L'estimateur des moments d'un n-échantillon gaussien de loi $\mathcal{N}(m, \sigma^2)$ pour le paramètre $\theta = (m, \sigma^2)$ est donné par le couple $(\overline{X}, S^2)$ (à faire en exerice).

2.4.1 Propriétés des estimateurs des moments

Soit $M(\theta) = \big(m_1(\theta), ..., m_p(\theta)\big)'$ le vecteur des moments d'ordre 1 à p de X et $\widehat{M}(\theta)$ celui des moments empiriques (les moments empiriques dépendent de θ via les v.a. X_i).

Propriété

- Si $M : \Theta \longrightarrow \mathbb{R}^p$ est bijective, alors

$$\widehat{\theta} = M^{-1} \circ \widehat{M}(\theta) \tag{72}$$

$$\widehat{M}(\theta) \underset{n \to +\infty}{\overset{\mathbb{P}_\theta\text{-p.s.}}{\longrightarrow}} M(\theta) \tag{73}$$

- Si M est bijective et M^{-1} est continue, alors

$$\widehat{\theta} \underset{n \to +\infty}{\overset{\mathbb{P}_\theta\text{-p.s.}}{\longrightarrow}} \theta \tag{74}$$

- Si de plus M^{-1} est différentiable et $\mathbb{P}_\theta$ a un moment d'ordre 2, alors

$$\sqrt{n}\big(\widehat{M}(\theta) - M(\theta)\big) \rightsquigarrow \mathcal{N}(0, \Sigma_\theta) \tag{75}$$

et

$$\sqrt{n}\big(\widehat{\theta} - \theta\big) \rightsquigarrow \mathcal{N}(0, \widetilde{\Sigma}_\theta) \tag{76}$$

avec $\Sigma_\theta = \big(\mathrm{cov}(X^i, X^j)\big)_{i,j=1..p}$ et $\widetilde{\Sigma}_\theta = J_M^{-1} \Sigma_\theta \big(J_M^{-1}\big)'$

Démonstration

La méthode des moments met en jeu des estimateurs plug-in dont les fonctionnelles statistiques sous-jacentes sont simplement les moments des v.a. X_i. Les propriétés de la méthode des moments découlent immédiatement des propriétés des estimateurs plug-in du paragraphe précédent:

- Si M est bijective, alors $\widehat{\theta}$ est défini de façon unique et la loi forte des grands nombres assure la convergence presque sûre de $\widehat{M}$ vers M (les X_i sont des v.a.i.i.d. et le moment d'ordre p existe par hypothèse).

- Si de plus M^{-1} est continue, alors le théorème de l'application continue montre que $\widehat{\theta}$ converge p.s. vers θ et l'on a donc consistance forte.

- Si M^{-1} est différentiable, le théorème de la limite centrale et la méthode delta s'appliquent (si $p > 1$).

J_M est la matrice jacobienne de M; en notant m la variable de M, le théorème d'inversion locale permet d'écrire que

$$J_{M^{-1}}(m) = J_M(\theta)^{-1} \tag{77}$$

□

Exemple:

Soit $(X_1, ..., X_n)$ un n-échantillon de $X \sim B(\alpha, \beta)$, dont la densité est

$$g(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}\mathbb{1}_{[0,1]}(x) \tag{78}$$

avec $\alpha, \beta > 0$. L'estimateur des moments de (α, β) est la solution de

$$\begin{cases} \overline{X} = \mathbb{E}[X] = \frac{\alpha}{\alpha+\beta} \\ \overline{X^2} = \mathbb{E}[X^2] = \frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)} \end{cases} \tag{79}$$

La fonction

$$\psi(x, y) = \left(\frac{x}{x+y}, \frac{x(x+1)}{(x+y)(x+y+1)} \right) \tag{80}$$

est de classe C^1 et inversible sur $]0,1[^2$. En inversant cette fonction, on trouve l'expression des estimateurs en fonction de $\overline{X}$ et $\overline{X^2}$:

$$\begin{cases} \widehat{\alpha} = \dfrac{\overline{X}(\overline{X^2} - \overline{X})}{\overline{X}^2 - \overline{X^2}} \\ \widehat{\beta} = \dfrac{(\overline{X^2} - \overline{X})(1-\overline{X})}{\overline{X}^2 - \overline{X^2}} \end{cases} \tag{81}$$

et ces estimateurs sont consistants et asymptotiquement normaux (à faire en exercice).

2.5 Estimateur du maximum de vraisemblance

2.5.1 Définition, propriétés et exemples

Définition 30

Un estimateur du maximum de vraisemblance de θ (on notera e.m.v.) est, sous réserve d'existence, une solution de l'équation

$$\widehat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmax}}\, L(X, \theta) \tag{82}$$

Ce maximum peut ne pas exister, n'être pas unique, ou être difficile à calculer.

Déterminer le maximum de vraisemblance est équivalent à déterminer le maximum de la log-vraisemblance (on notera cette log-vraisemblance $l(X,\theta)$). Si L est différentiable, l'estimateur du maximum de vraisemblance est alors solution de l'équation

$$\frac{\partial}{\partial\theta} \ln L(X,\theta) = \frac{\partial}{\partial\theta} l(X,\theta) = \nabla l(X,\theta) = S(X,\theta) = 0 \tag{83}$$

La dérivée S de la log-vraisemblance (ou son gradient dans le cas multidimensionnel) s'appelle le score.

L'équation précédente donne une condition nécessaire, mais non suffisante, pour être un maximum local. On rappelle qu'un maximum est local s'il existe un voisinage de $\widehat{\theta}$ dans Θ tel que

$$L(x,\theta) < L(x,\widehat{\theta}) \tag{84}$$

pour tout θ dans ce voisinage. Un maximum est global si l'inégalité précédente est vérifiée pour tout $\theta \in \Theta$.

Dans le cas où Θ est un intervalle de $\mathbb{R}$ (le paramètre à estimer est scalaire), une condition suffisante d'existence et d'unicité est que $L(x,\theta)$ soit une fonction de θ de classe C^2 sur Θ (x étant fixé) et que les deux conditions suivantes soient réalisées:

$$\begin{cases} \frac{\partial l}{\partial \theta}(x,\theta) = 0 \\ \\ \frac{\partial^2 l}{\partial \theta^2}(x,\theta) < 0 \end{cases} \tag{85}$$

La solution doit par ailleurs appartenir à l'intérieur $\overset{\circ}{\Theta}$ de Θ. Les conditions assurent l'existence et l'unicité d'un maximum local uniquement et la valeur $\widehat{\theta}$ qui réalise ce maximum n'est donc pas forcément l'estimateur du maximum de vraisemblance (car il peut ne pas être un maximum global).

On a existence et unicité d'un maximum global si la fonction l est strictement concave sur Θ (en tant que fonction de Θ).

Le cas multidimensionnel va être abordé après les quelques exemples qui suivent.

Exemple:

Soit $X = (X_1, ..., X_n)$ un échantillon gaussien $\mathcal{N}(m, \sigma^2)$ et $\theta = (m, \sigma^2)$.

Le modèle est paramétrique, dominé par la mesure de Lebesgue et la vraisemblance associée est

$$L(x,\theta) = \prod_{i=1}^{n} f(x_i,\theta) = (2\pi)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - m)^2 - \frac{n}{2}\ln\sigma^2\right) \tag{86}$$

Le support du modèle est $\mathbb{R}^n$; celui-ci est donc homogène et la vraisemblance est clairement de classe C^2 en θ (à condition que $\sigma > 0$). Par conséquent, l'e.m.v. est solution de l'équation de vraisemblance

$$\frac{d}{d\theta} l(X,\theta) = \nabla_l(\theta) = 0 \tag{87}$$

où $l(x,\theta) = \ln L(x,\theta)$. La solution est unique si la matrice hessienne $H_l(\widehat{\theta})$ est définie négative.

$$l(x,\theta) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - m)^2 - \frac{n}{2}\ln\sigma^2 - \frac{n}{2}\ln(2\pi) \tag{88}$$

$l(x,\theta)$ est une fonction vectorielle en θ, car $\theta = (m,\sigma^2)$ est de dimension 2. La dérivée de l est donc à comprendre ici comme le gradient de cette fonction, dont les coordonnées sont les dérivées partielles en m et en σ^2.

$$\frac{d}{d\theta} l(x,\theta) = \nabla_l(\theta) = \left(\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - m); \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i - m)^2 - \frac{n}{2\sigma^2}\right)' \tag{89}$$

Ce gradient s'annule si, et seulement si

$$\sum_{i=1}^{n}(x_i - m) = 0 \text{ et } \sum_{i=1}^{n}(x_i - m)^2 = n\sigma^2 \tag{90}$$

$$\iff m = \overline{x} \text{ et } \sigma^2 = S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 \tag{91}$$

Un point critique de l est donc $\widehat{\theta} = \left(\overline{X}, S^2\right)$. La matrice hessienne de l en ce point est (à vérifier à titre d'exercice)

$$H_l(\widehat{\theta}) = \begin{pmatrix} -n/S^2 & 0 \\ 0 & -n/(2S^4) \end{pmatrix} \tag{92}$$

qui est clairement définie négative. Par suite, $\widehat{\theta}$ est l'unique e.m.v. Pour vérifier les calculs précédents, on aura intérêt à poser $S^2 = V$ pour éviter les erreurs au moment de la dérivations en S^2.

Exemple:

Soit $X = (X_1, ..., X_n)$ un n-échantillon de loi uniforme sur $[0,\theta]$ où $\theta > 0$.

Le modèle est paramétrique, dominé par la mesure de Lebesgue sur $\mathbb{R}^n$. La vraisemblance de l'échantillon est

$$L(x,\theta) = \prod_{i=1}^{n}\left(\frac{1}{\theta}\mathbb{1}_{[0,\theta]}(x_i)\right) = \frac{1}{\theta^n}\mathbb{1}_{[0,\theta]^n}(x) = \frac{1}{\theta^n}\mathbb{1}[0 \le x_{(1)} \le x_{(n)} \le \theta](x) \tag{93}$$

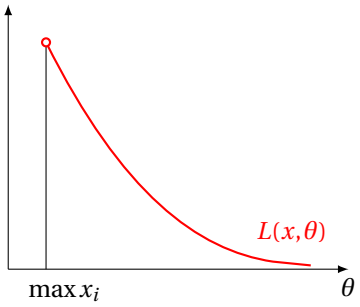

Figure 2.5 – Vraisemblance d'un échantillon uniforme

où $x_{(1)}$ est le min des x_i et $x_{(n)}$ le max. Le support dépend de θ et le modèle n'est donc pas homogène. Nous ne pouvons pas déterminer le maximum de vraisemblance par la méthode précédente. On remarque par contre que, à x fixé, la fonction de vraisemblance est décroissante en θ à partir de $x_{(n)}$ (avant, elle est nulle). Le maximum est donc atteint en $\widehat{\theta} = x_{(n)}$. L'e.m.v. existe et est unique:

$$\widehat{\theta}_{MV} = X_{(n)} = \max_{i=1}^{n} X_i \tag{94}$$

On montre facilement que $\theta X_{(n)}$ suit une loi béta de paramètres $(n,1)$. La densité du maximum est donc

$$\frac{n y^{n-1}}{\theta} \mathbb{1}_{[0,\theta]}(y) \tag{95}$$

Ainsi, $\mathbb{E}[X_{(n)}] = \frac{n}{n+1}\theta$ et l'e.m.v. a donc un biais égal à $b(\theta) = -\frac{\theta}{n+1}$.

Exemple:

Si $X_1, ..., X_n$ est un n-échantillon de loi uniforme sur $[\theta, \theta+1]$, on peut montrer (en exercice) que tout estimateur de θ compris entre $X_{(n)} - 1$ et $X_{(1)}$ est un estimateur du maximum de vraisemblance. Il n'y a alors pas unicité de cet e.m.v.

Exemple: Familles exponentielles

Considérons une famille exponentielle canonique de paramètre $\theta_1, ..., \theta_p$, de statistique canonique $T = (T_1, ..., T_p)$ et de fonction de log-partition $\beta(\theta)$.

Si β est de classe C^2 sur Θ et possède une matrice hessienne définie négative alors l'e.m.v. existe, est unique, et vérifie

$$\frac{\partial}{\partial \theta_i} \beta\left(\widehat{\theta}\right) = -\frac{1}{n} \sum_{j=1}^{n} T_i(X_j) \tag{96}$$

L'e.m.v. peut ne pas exister, par exemple dans le cas où la vraisemblance L n'est pas une fonction continue de θ. Pour avoir un estimateur intéressant, il faudra donc *a minima* supposer la vraisemblance continue, voire différentiable et également imposer à Θ certaines propriétés :

Théorème 31 — Condition suffisante d'existence de l'e.m.v.

Si Θ est un espace compact et si l est continue sur Θ, alors l'estimateur du maximum de vraisemblance existe.

Théorème 32

Sous les trois hypothèses de régularité ($\star$) suivantes:

- Le modèle est identifiable et homogène.
- Θ est un ouvert ou un compact de $\mathbb{R}^p$.
- Pour tout x, la fonction $L(x,\theta)$ est de classe C^2 en θ sur Θ.

Alors l'estimateur du maximum de vraisemblance $\widehat{\theta}$ est solution $\mathbb{P}_\theta$-p.s. du système

$$\begin{cases} \nabla l(x,\theta) = 0 \\ H_l(x,\theta) < 0 \end{cases} \tag{97}$$

où $l(x,\theta) = \ln L(x,\theta)$ est la log-vraisemblance de l'échantillon, ∇l le gradient de l en θ (x est fixé) et H_l est la matrice hessienne de l en θ (à x fixé).

La condition $H_l(x,\theta) < 0$ signifie que la matrice hessienne est définie négative en θ. Le théorème n'assure toujours pas l'unicité de l'e.m.v. (il est unique localement). Une condition suffisante pour garantir l'unicité est de s'assurer que Θ est un espace convexe et que l est strictement concave sur Θ.

2.5.2 Reparamétrisation

Trouver un estimateur par substitution de l'estimateur du maximum de vraisemblance s'appelle une reparamétrisation. On montre que quelque soit l'application ψ, l'e.m.v. est invariant par reparamétrisation.

Théorème 33 de Zehna

Soit $\widehat{\theta}$ l'e.m.v. de θ et ψ une fonction quelconque. Alors l'e.m.v. de $\psi(\theta)$ est $\psi(\widehat{\theta})$

Démonstration :

Soit $\eta = \psi(\theta)$. Si ψ est bijective, on a immédiatement

$$\widetilde{L}(x,\eta) = L(x,\psi^{-1}(\eta)) \tag{98}$$

Dans le cas contraire,

$$\widetilde{L}(x,\eta) = \sup_{\theta/\psi(\theta)=\eta} L(x,\theta) \tag{99}$$

Soit $\widehat{\eta}$ l'e.m.v. de η. Alors par définition, $\widehat{\eta}$ maximise la vraisemblance:

$$\widetilde{L}(x,\widehat{\eta}) = \sup_{\eta} \widetilde{L}(x,\eta) = \sup_{\eta} \sup_{\theta/\psi(\theta)=\eta} L(x,\theta) = L(x,\widehat{\theta}) = \widetilde{L}(x,\psi(\widehat{\theta})) \tag{100}$$

Par identification, on a alors $\widehat{\eta} = \psi(\widehat{\theta})$.

□

2.5.3 Propriétés asymptotiques

C'est Fisher qui a le premier étudié, en 1921 et 1925, l'estimateur du maximum de vraisemblance. Il a alors défini les notions d'efficacité et d'exhaustivité qui seront étudiées dans le chapitre suivant. En 1946, Cramer a donné une première preuve de la consistance de cet estimateur et de sa normalité asymptotique, sous certaines conditions de régularité. Wald a démontré la consistance forte en 1949, sous des hypothèses plus générale. Rappelons que $S = \partial l/\partial\theta$ s'appelle le score du modèle.

Théorème 34 de Cramer (1946)

On considère le modèle d'échantillonnage $(\mathbb{P}_\theta)_{\theta\in\Theta}$ de vraisemblance $L(X,\theta) = L(X_1,...,X_n,\theta)$ absolument continue par rapport à $\mathbb{P}_\theta$.

Sous les hypothèses de régularité ci-dessous,

- H0 : Θ est un ouvert ou un compact de $\mathbb{R}$.
- H1 : Pour presque tout x, pour tout $i = 0...2$, $\dfrac{\partial^i l}{\partial\,\theta^i}$ existe $\forall\ \theta\in\Theta$
- H2 : Pour presque tout x, pour tout $i = 0...2$, $\left|\dfrac{\partial^i l}{\partial\,\theta^i}(x,\theta)\right| < F_i(x)\ \forall\ \theta\in\Theta$

où les F_i sont intégrables sur $\mathbb{R}$ et $\displaystyle\int_{\mathbb{R}} F_2(x)L(x,\theta)dx < M$ qui ne dépend pas de θ.

- H3 : Pour tout $\theta\in\Theta$, $0 < \mathbb{E}_\theta\left[S(X,\theta)^2\right] < \infty$

alors il existe une suite $\left(\widehat{\theta}_n\right)_n$ de solutions de l'équation de vraisemblance $S(X,\theta) = 0$ qui converge en probabilité vers la vraie valeur du paramètre θ lorsque le nombre d'observations n tend vers l'infini.

Le théorème démontre, dans le cas unidimensionnel où $\Theta\subset\mathbb{P}$, la consistance faible de certaines solutions de l'équation de vraisemblance (qui ne sont donc pas forcément égales au maximum absolu de vraisemblance) vers la vraie valeur de θ.

Démonstration : Théorème admis (cf. article de Cramer). □

Il existe une foultitude de conditions de régularité différentes selon les ouvrages, qui conduisent à la consistance forte ou faible de l'e.m.v. Voici un énoncé dans le cadre multidimensionnel, avec des hypothèses de régularité que nous adopterons par défaut et qui sont suffisamment pratiques pour couvrir la plupart des cas classiquement étudiés.

Théorème 35 Consistance forte de l'e.m.v.

On considère le modèle d'échantillonnage $X_1,...,X_n$ de v.a.i.i.d. $\sim L(x,\theta)$. Sous les hypothèses de régularité $(\star\star)$ ci-dessous,

- Le modèle est identifiable et homogène.
- Θ est un ouvert ou un compact de $\mathbb{R}^p$.
- Pour tout x, la fonction $L(x,\theta)$ est différentiable en θ sur Θ.

Alors l'estimateur du maximum de vraisemblance est fortement consistant.

Démonstration :

Soit $\widehat{\theta}_n = \widehat{\theta}(X_1,...,X_n)$ l'e.m.v. de θ, dans un modèle d'échantillonnage $(X_1,...,X_n)$ vérifiant les propriétés de régularité $(\star\star)$. $\widehat{\theta}_n$ est une statistique fonction des observations $X_1,...,X_n$. On suppose que les X_i suivent la loi d'une v.a. X dont la densité est $L(x,\theta_0)$ avec $\theta_0\in\Theta$. $L_n(\theta) = L_n(x_1,...,x_n,\theta)$ est la vraisemblance de l'échantillon.

Nous voulons démontrer que

$$\mathbb{P}_\theta\left[\lim_{n\to+\infty}\widehat{\theta}_n = \theta\right] = 1 \tag{101}$$

$$\lim_{n\to+\infty}\widehat{\theta}_n(\omega) = \theta \iff \forall\epsilon>0, \exists\ N\in\mathbb{N}/\ \forall n\geq N, \left|\widehat{\theta}_n - \theta\right| < \epsilon \tag{102}$$

ainsi,

$$\left[\lim_{n\to+\infty}\widehat{\theta}_n = \theta\right] = \bigcap_{\epsilon>0}\bigcup_{N\in\mathbb{N}}\bigcap_{n\geq N}\left[\left|\widehat{\theta}_n - \theta\right| < \epsilon\right] = \bigcap_{\epsilon>0} B_\epsilon = \bigcap_{\epsilon>0}\bigcup_{N\in\mathbb{N}} B_N(\epsilon) \tag{103}$$

Pour $1/(p+1) < \epsilon < 1/p$, on a $B_{1/(p+1)}\subset B_\epsilon\subset B_{1/p}$, de sorte que

$$\left[\lim_{n\to+\infty}\widehat{\theta}_n = \theta\right] = \bigcap_{p\in\mathbb{N}^*} B_{1/p} \tag{104}$$

où $\left(B_{1/p}\right)_{p>0}$ est une suite dénombrable décroissante d'évènements. Ainsi, si nous arrivons à démontrer que

$$\forall \epsilon > 0,\ \mathbb{P}_\theta(B_\epsilon) = 1 \tag{105}$$

alors par intersection dénombrable,

$$\mathbb{P}_\theta\left(\bigcap_{p>0} B_{1/p}\right) = 1 \tag{106}$$

et sur cet évenèment presque sûr, $\lim_n \widehat{\theta}_n(\omega) \longrightarrow \theta$. Posons maintenant

$$\Lambda_n(\theta) = \frac{1}{n} LLR(\theta,\theta_0) = \frac{1}{n}\ln\frac{L_n(\theta)}{L_n(\theta_0)} = \frac{1}{n}\sum_{i=1}^{n}\ln\frac{L(X_i,\theta)}{L(X_i,\theta_0)} \tag{107}$$

La loi forte des grands nombres s'applique et $\Lambda_n(\theta)$ converge $\mathbb{P}_\theta$-p.s. vers

$$\mathbb{E}_\theta\left[\ln\frac{L(X,\theta)}{L(X,\theta_0)}\right] = -K(\theta_0,\theta) \tag{108}$$

Par définition de θ_0, $-K(\theta,\theta_0) \leq 0$ et cette fonction est nulle si, et seulement si, $L(X,\theta) = L(X,\theta_0)$ avec probabilité 1. Par identifiabilité du modèle, cela implique $\theta = \theta_0$ et pour tout $\theta \neq \theta_0$, il existe donc $N_\epsilon \in \mathbb{N}$ tel que

$$\mathbb{P}_{\theta_0}\left(\cap_{n>N_\epsilon}[\Lambda_n(\theta) < 0]\right) = 1 \tag{109}$$

Pour tout ϵ suffisamment petit de telle sorte que $[\theta_0-\epsilon,\theta_0+\epsilon] \subset \Theta$, il existe un entier N_ϵ tel que

$$\mathbb{P}_{\theta_0}\left(\cap_{n>N_\epsilon}[\Lambda_n(\theta_0 \pm \epsilon) < 0]\right) = 1 \tag{110}$$

La fonction Λ_n est continue et atteint son maximum dans l'intervalle compact $[\theta_0-\epsilon,\theta_0+\epsilon]$. Soit $\widehat{\theta}_n$ le point le plus proche de θ_0 pour lequel ce maximum est atteint. Par définition, $\Lambda_n(\widehat{\theta}_n) \geq \Lambda_n(\widehat{\theta}_0) = 0$. Le maximum est donc réalisé à l'intérieur de l'intervalle car $\Lambda_n(\theta_0 \pm \epsilon) < 0$. $\forall \epsilon > 0$ assez petit, il existe N_ϵ tel que

$$\mathbb{P}_{\theta_0}\left(\bigcap_{n>N_\epsilon}[\exists \theta_n \text{ solution de l'équation de vraisemblance }]\right) = 1 \tag{111}$$

□

Parmi d'autres énoncés, on peut retenir (sans démonstration) le théorème suivant

Théorème 36 Gouriéroux & Monfort, propriété 7.10. « statistiques et modèles économétriques »

On considère un n-échantillon $X_1,...,X_n$ de $X \sim L(x,\theta)$. On note $l_n(X,\theta)$ la fonction de log-vraisemblance de l'échantillon. Sous les hypothèses de régularité (••) ci-dessous,

- Le modèle est localement identifiable en θ_0, vraie valeur du paramètre.
- Θ est un compact de $\mathbb{R}^p$.
- Pour tout x, la fonction $l_n(x,\theta)$ est continue en θ sur Θ.
- $l_n(X,\theta)/n$ converge presque sûrement vers $\mathbb{E}_{\theta_0}[l(X,\theta)] < \infty$ uniformément en $\theta \in \Theta$.

Alors il existe une suite d'e.m.v. convergeant presque sûrement vers θ_0.

Si l'intérieur de Θ n'est pas vide et que $\theta_0 \in \overset{o}{\Theta}$, l'hypothèse de compacité de Θ n'est pas nécessaire.

Si la fonction de vraisemblance est différentiable, alors on peut remplacer « il existe une suite d'e.m.v.» par «il existe une suite de solutions des équations de vraisemblance ».

2.6 Estimation dans un cadre bayésien

On rappelle que dans un cadre bayésien, le modèle statistique possède un paramètre muni d'une loi de probabilité *a priori*. Cette loi mesure l'information sur θ disponible avant la collecte et la prise en compte des observations. L'inférence bayésienne est construite sur la loi *a posteriori* qui mesure la mise à jour de l'information sur θ au fur et à mesure de la prise en compte des observations. Le calcul de la loi *a posteriori* s'effectue à partir de la formule de Bayes:

$$L(\theta|x) \propto L(x|\theta) \times \Pi(\theta) \tag{112}$$

Le symbole $\propto$ signifie « proportionnel à ». T est la variable aléatoire sous-jacente à θ et sa loi est donc Π.

Voici plusieurs exemples d'estimateurs bayésiens:

Moyenne *a posteriori*

$$\mathbb{E}[T|X=x] = \int_{\Theta} \theta \times L(\theta|x)d\theta \tag{113}$$

Médiane *a posteriori*

$$\mathbb{P}[T \leq \theta_{me}] = \int_{-\infty}^{\theta_{me}} L(\theta|x)d\theta = 1/2 \tag{114}$$

Mode *a posteriori*

$$\underset{\theta}{\operatorname{argmax}}\, L(\theta|x) \tag{115}$$

Exemple:

Nous reprenons l'exemple de la fin du chapitre précédent; $(X_1,...,X_n)$ est un n-échantillon de $X \sim \mathcal{N}(\theta,\sigma^2)$ dans lequel le paramètre inconnu θ suit une loi *a priori* normale $\mathcal{N}(a,b^2)$. σ^2, a et b sont supposés connus. L'*a priori* est informatif si b est petit, c'est à dire si l'on est sûr que θ est voisin de a. On rappelle que la loi *a posteriori* du modèle est

$$\mathcal{N}\left(\frac{\overline{x}b^2 + a\sigma^2/n}{b^2+\sigma^2/n}; \frac{b^2\sigma^2/n}{b^2+\sigma^2/n}\right) \tag{116}$$

L'estimateur de θ pour la moyenne *a posteriori* est

$$\widehat{\theta} = \overline{X}\frac{b^2}{b^2+\sigma^2/n} + a\left(1 - \frac{b^2}{b^2+\sigma^2/n}\right) \tag{117}$$

c'est une moyenne pondérée de la moyenne empirique $\overline{x}$ et de la moyenne *a priori* a.

Si $n \to +\infty$, $\widehat{\theta} \longrightarrow \overline{X}$. Les approches fréquentistes et bayésiennes coïncident.

Sinon, si l'*a priori* est informatif (b petit), alors $\widehat{\theta} \approx a$. Les observations ne sont pratiquement pas prises en compte. Si par contre l'*a priori* n'est pas informatif (b grand), alors la loi *a priori* n'apporte pratiquement pas d'information. Toute l'information sur θ est comprise dans $\overline{X}$.

Exemple:

Reprenons maintenant le second exemple de la fin du chapitre précédent ; considérons un n-échantillon $X = (X_1,...,X_n)$ de loi de Bernoulli de paramètre θ inconnu. Nous avons vu que la vraisemblance de l'échantillon est

$$L(x,\theta) = \prod_{i=1}^{n} L(x_i,\theta) = \theta^s(1-\theta)^{n-s} \tag{118}$$

avec $s = \sum_{i=1}^{n} x_i$, et si la loi *a priori* est une loi béta, alors la vraisemblance *a posteriori* suit également une loi béta:

$$L(\theta|x) \propto \theta^s(1-\theta)^{n-s}\frac{1}{B(a,b)}\theta^{a-1}(1-\theta)^{b-1}\mathbb{1}_{[0,1]}(\theta) \tag{119}$$

Une estimation possible de θ peut être donnée par la moyenne *a posteriori* de la loi béta:

$$\widehat{\theta} = \mathbb{E}[\theta|X] = \frac{S+a}{(S+a)+(n-S+b)} \tag{120}$$

Chapitre 3

Estimation optimale

3.1 Exhaustivité, minimalité, complétude

3.1.1 Exhaustivité

Exemple:

$X = (X_1, ..., X_n)$ n-échantillon de loi de bernoulli de paramètre $\theta \in]0,1[$. Soit $S = X_1 + X_2 + ... + X_n$ une statistique de cet échantillon.

L'espace des obervations est $\mathbb{X} = \{x = (x_1, ..., x_n) : x_i = 0,1\}$, de cardinal 2^n. On peut partitionner cet espace en n évènements à l'aide de S:

$$\mathbb{X} = \bigcup_{i=0}^{n} [S = i] \tag{1}$$

S suit une loi binomiale de paramètres n et θ. Notons $[X = x] = \bigcap_{i=1}^{n} [X_i = x_i]$

$$\mathbb{P}[X_i = x_i] = \theta^{x_i}(1-\theta)^{1-x_i}\mathbb{1}_{\{0,1\}}(x_i) \tag{2}$$

$$\mathbb{P}[X = x | S = s] = \frac{\mathbb{P}([X = x] \cap [S = s])}{\mathbb{P}[S = s]} = \frac{\prod_{i=1}^{n}\left(\theta^{x_i}(1-\theta)^{1-x_i}\right)}{\begin{pmatrix} n \\ s \end{pmatrix}\theta^s(1-\theta)^{n-s}}\mathbb{1}_{B_s}(x) = \frac{1}{\begin{pmatrix} n \\ s \end{pmatrix}}\mathbb{1}_{B_s}(x) \tag{3}$$

avec $B_s = \{(x_1, ..., x_n) : \sum_{i=1}^{n} x_i = s\}$.

Cette expression ne dépend pas de θ: le fait de savoir la valeur de l'observation x lorsque l'on connaît la somme n'apporte pas plus d'information sur θ. Autrement dit, S condense l'information apportée par le vecteur $(X_1, ..., X_n)$ sans la dégrader. Les données originelles de l'échantillon ne contiennent pas d'information supplémentaire sur la loi de X et peuvent être écartées. S suffit pour apporter toute l'information (en anglais, le terme correspondant à « exhaustif » est « sufficient »).

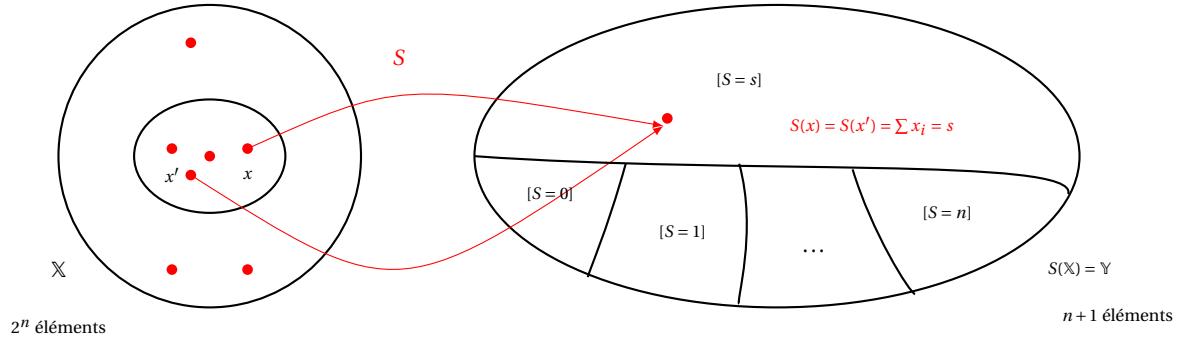

Figure 3.1 – Exhaustivité et tribu engendrée

Soit $(\mathbb{X}, \mathfrak{F}, (\mathbb{P}_\theta)_{\theta \in \Theta})$ un modèle paramétrique dominé.

Définition 37

Une statistique S est exhaustive pour X si la loi conditionnelle de X sachant $[S = s]$ ne dépend pas de θ.

Lorsque la valeur prise par une statistique exhaustive S est connue et égale à s, alors l'échantillon ne fournit plus d'information sur le paramètre inconnu θ car sa loi ne dépend plus de θ. La statistique S condense alors parfaitement la totalité de l'information de l'échantillon, relativement au paramètre θ. Elle contient toute l'information nécessaire à l'inférence sur le paramètre, sans la dégrader.

Une autre façon de comprendre la notion d'exhaustivité est de considérer une observation x au travers d'une statistique $S(x)$. L'information sur θ sera la même pour une autre observation y si la statistique S est exhaustive et que $S(x) = S(y)$. La connaissance de S suffit à obtenir toute l'information sur θ.

Théorème 38 Critère de factorisation de Neyman

S est exhaustive si, et seulement si, la vraisemblance du modèle s'écrit

$$L(x,\theta) = h(x) \times g(S(x),\theta) \tag{4}$$

Démonstration dans le cas discret:

Nous commençons par exposer le cas des v.a. discrètes, plus simple que le cas des v.a. continues. Supposons donc S exhaustive. Alors la densité conditionnelle de X sachant $[S = s]$ ne dépend pas de θ. Posons $s = S(x)$ pour un x fixé.

$$L(x,\theta) = \mathbb{P}_\theta[X = x] = \mathbb{P}_\theta[X = x|S = s]\mathbb{P}_\theta[S = s] = h(x) \times g(s,\theta) \tag{5}$$

Réciproquement, supposons que $L(x,\theta) = h(x)g(s,\theta)$

$$\mathbb{P}_\theta[S = s] = \mathbb{P}_\theta[S(x) = s] = \sum_{y\in\mathbb{X}} \mathbb{1}_{[S(y)=s]}\mathbb{P}_\theta[X = y] = \sum_{y\in\mathbb{X}} \mathbb{1}_{[S(y)=s]}h(y)g(S(y),\theta) \tag{6}$$

$$= g(t,\theta)\sum_{y\in\mathbb{X}} \mathbb{1}_{[S(y)=s]}h(y) \tag{7}$$

ainsi,

$$\mathbb{P}_\theta[X = x, S = s] = \mathbb{P}_\theta[X = x, S(x) = s] = \begin{cases} 0 \text{ si } S(x) \neq t \\ \mathbb{P}_\theta[X = x] \text{ si } S(x) = s \end{cases} \tag{8}$$

$$= \mathbb{1}_{[S(x)=s]}g(t,\theta)h(x) \tag{9}$$

On a donc bien

$$\mathbb{P}_\theta[X = x|S = s] = \frac{\mathbb{1}_{h(x)[S(x)=s]}}{\sum_{y\in\mathbb{X}} \mathbb{1}_{[S(y)=s]}h(y)} \tag{10}$$

qui ne dépend pas de θ. S est donc exhaustive.

□

La démonstration générale (initialement due à Jerzy Neyman) utilise les deux lemmes donnés ci-dessous sans démonstration. Ces lemmes sont des résultats de théorie de la mesure et sont démontrés, par exemple, dans « Mathematical Statistics » de Jun Shao (Springer).

Lemme

Soit $(a_n)_n$ une suite de réels strictement positifs tels que $\sum_{n=1}^{\infty} a_n = 1$. Soit $(P_n)_n$ une suite de mesures de probabilités sur un même espace et soit $P = \sum_{n=1}^{\infty} a_n P_n$.

Alors P est une mesure de probabilité et $P \ll \nu \iff P_n \ll \nu,\ \forall n$. De plus, si ν est σ-finie,

$$\frac{dP}{d\nu} = \sum_{n=1}^{\infty} a_n \frac{dP_n}{d\nu} \tag{11}$$

Lemme

Si la famille $\mathscr{P}$ est dominée par une mesure σ-finie ν, alors $\mathscr{P}$ est également dominée par $Q = \sum_{n=1}^{\infty} a_n P_n$ où $P_n \in \mathscr{P}$ et $(a_n)_n$ suite de réels positifs dont la somme vaut 1.

Démonstration dans le cas continu:

Supposons T exhaustive pour une mesure $\mathbb{P} \in \mathscr{P}$, où $\mathscr{P}$ est une famille de probabilités sur $(\mathbb{R}^n, B(\mathbb{R}^n))$. $\mathbb{P}(A|T)$ ne dépend pas de la mesure P. Soit A un borélien de $\mathbb{R}^n$ et soit $Q = \sum_{i=1}^{\infty} a_i P_i$. D'après le lemme 1 et d'après le théorème de Fubini, on a:

$$Q(A \cap B) = \sum_{i \geq 1} a_i P_i(A \cap B) = \sum_i a_i \int_B P(A|T) dP_i = \int_B P(A|T) dQ \tag{12}$$

où B appartient à la tribu engendrée par T. Ainsi, $\mathbb{P}(A|T) = \mathbb{E}_Q[\mathbb{1}_A|T]$ Q-p.s. Notons $g(T)$ la dérivée au sens de Radon-Nikodyn de P par rapport à Q, sur l'espace $(\mathbb{R}^n, \sigma(T), Q)$: $g(T) = dP/dQ$. On a

$$P(A) = \int P(A|T) dP = \int \mathbb{E}_Q[\mathbb{1}_A|T] g(T) dQ = \int \mathbb{E}_Q[\mathbb{1}_A g(T)|T] dQ = \int_A g(T) \frac{dQ}{d\nu} d\nu \tag{13}$$

Par suite, on a bien $L(x,\theta) = h(x) g(T,\theta)$ en posant $h = dQ/d\nu$.

Réciproquement, si $dP/d\nu = L(x,\theta) = g(T(x),\theta) h(x)$. Alors:

$$dP/dQ = dP/d\nu \div \left(\sum_i a_i dP_i/d\nu \right) = g(T)/\sum_i g_{P_i}(T) \tag{14}$$

d'après le lemme 2. Soit alors $A \in \sigma(X)$ et $P \in \mathscr{P}$. Par exhaustivité de T, on a

$$P(A|T) = \mathbb{E}_Q[\mathbb{1}_A|T] \tag{15}$$

Par définition de l'espérance conditionnelle, on a alors

$$\int_B \mathbb{1}_A dP = \int B \mathbb{E}_Q[\mathbb{1}_A|T] dP \tag{16}$$

Pour tout $B \in \sigma(T)$. Par ailleurs, dP/dQ est une fonction borélienne de T. On a donc

$$\int_B \mathbb{1}_A dP = \int B \mathbb{E}_Q[\mathbb{1}_A|T] dP = \int_B \mathbb{E}_Q\left[\mathbb{1}_A \frac{dP}{dQ} | T \right] dQ = \int_B \mathbb{1}_A \frac{dP}{dQ} dQ \tag{17}$$

□

Exemple:

Dans le modèle d'échantillonnage de Bernoulli du début de la leçon, la vraisemblance s'écrit

$$L(x,\theta) = \theta^{S(x)} (1-\theta)^{n-S(x)} = h(x) \times g(s,\theta) \tag{18}$$

On a la décomposition avec $h(x) = 1$, $S(x) = \sum_{i=1}^n x_i$ et $g(s,\theta) = \theta^s (1-\theta)^{n-s}$; S est donc exhaustive.

Exemple:

Considérons un n-échantillon $X = (X_1, ..., X_n)$ de loi uniforme sur $[\theta - 1/2, \theta + 1/2]$. Le modèle est dominé par la mesure de Lebesgue (mais pas homogène) et sa vraisemblance s'écrit

$$L(x,\theta) = \prod_{i=1}^n \mathbb{1}_{[\theta-1/2,\theta+1/2]}(x_i) = \mathbb{1}_{[\theta - 1/2 \leq x_{(1)} \leq x_{(n)} \leq \theta + 1/2]}(x) = \mathbb{1}_{J(\theta)}(x) \tag{19}$$

que l'on peut écrire sous la forme $g((x_{(1)}, x_{(n)}), \theta) \times 1$ de sorte que $(X_{(1)}, X_{(n)})$ est une statistique exhaustive. Par contre l'une de ces deux statistiques ne serait pas exhaustive seule.

Exemple:

De façon générale, pour un n-échantillon $X = (X_1, ..., X_n)$ de loi quelconque, la vraisemblance est

$$L(x,\theta) = \prod_{i=1}^n f(x_i) = \prod_{i=1}^n f(x_{(i)}) \tag{20}$$

et par suite, la statistique d'ordre $X = (X_{(1)}, ..., X_{(n)})$ est exhaustive. Elle ne condense par contre pas du tout l'information...

Exemple: Modèle exponentiel

Dans un modèle exponentiel, la vraisemblance s'écrit

$$L(x,\theta) = h(x) \times g(t,\theta) = h(x) \exp\left(\langle T(x), \lambda(\theta)\rangle - \beta(\theta)\right) \tag{21}$$

et donc la statistique naturelle est exhaustive. En particulier, la statistique canonique est exhaustive.

C'est le cas par exemple d'un n-échantillon gaussien de loi $\mathcal{N}(m,\sigma^2)$ avec $\theta = (m,\sigma^2)$. La statistique $T(X) = (T_1(X), T_2(X)) = (\overline{X}, S^2)$ est exhaustive pour θ et la décomposition de la vraisemblance correspondante est donnée par

$$L(x,\theta) = (2\pi)^{-n/2} \exp\left(-\frac{n}{2\sigma^2}\left[(t_1 - m)^2 + t_2\right] - \frac{n}{2}\ln(\sigma^2)\right) \tag{22}$$

avec $h(x) = (2\pi)^{-n/2}$, $t = (t_1, t_2) = (\overline{x}, s^2)$.

Attention, dans cet exemple, c'est le théorème de factorisation qui permet de conclure, car l'expression à l'intérieur de la parenthèse n'est pas sous la forme d'un produit scalaire (on pourrait la transformer pour faire apparaître ce produit scalaire).

Théorème 41

Soit T une statistique exhaustive et ϕ une fonction mesurable. Alors la statistique S vérifiant $T = \phi(S)$ est exhaustive.

Si ϕ est bijective, S et T sont alors équivalentes et T exhaustive $\iff$ S exhaustive.

Autrement dit, l'image réciproque d'une statistique exhaustive est exhaustive. L'existence d'une fonction ϕ telle que $T = \phi(S)$ induit une relation d'ordre partiel sur les statistiques. On notera $T < S$. $T < S$ et T exhaustive entraîne S exhaustive, car S est plus fine que T.

Démonstration:

Si toutes les mesures de probabilités en jeu sont dominées par une mesure σ-finie ν, alors il suffit d'écrire la vraisemblance:

$$L(x,\theta) = g(t,\theta)h(x) = g(\phi(s),\theta)h(x) = \tilde{g}(s,\theta)h(x) \tag{23}$$

De façon plus générale, si S n'est pas exhaustive pour une mesure P, alors il existe deux mesures P_1 et P_2 telles que les distributions conditionnelles de X sachant $S(X)$ sous P_1 et P_2 soient différentes. Considérons la sous-famille $\mathscr{P}_0$ de $\mathscr{P}$. $T(X)$ est exhaustive pour $\mathscr{P}$ donc aussi pour $\mathscr{P}_0$. Mais la mesure $P_1 + P_2$ domine $\mathscr{P}_0$ et donc $S(X)$ est exhaustive pour P dans $\mathscr{P}_0$. Par suite, les distributions conditionnelles de S sachant $S(X)$ sont les mêmes sous P_1 et P_2, ce qui est absurde.

□

Si ϕ n'est pas bijective, alors la tribu induite va être plus grossière : $\sigma(T) \subset \sigma(S)$ et l'information va être plus condensée. On peut se demander s'il existe une statistique qui condense l'information de façon maximale sans la dégrader.

Définition 42

Une statistique exhaustive S est minimale si elle est fonction de toute autre statistique exhaustive. Autrement dit:

$$T \text{ exhaustive minimale} \iff \forall S \text{ exhaustive}, \exists\, \phi :\ T = \phi(S),\ \mathbb{P}_\theta - \text{p.s.}, \forall\theta \tag{24}$$

Une statistique T est fonction d'une statistique S si, et seulement si, $S(x) = S(x') \Rightarrow T(x) = T(x')$. Autrement dit, les évènements de la forme $[S(x) = y']$ sont chacun inclus dans un évènement de la forme $[T(x) = y]$. La partition associée à une statistique exhaustive minimale est donc la plus grossière possible et la réduction est maximale.

On définit le rapport de vraisemblance (pour « likelihood ratio ») par la formule suivante, pour x, y fixés:

$$\theta \longrightarrow LR(x,y,\theta) = \frac{L(x,\theta)}{L(y,\theta)} \tag{25}$$

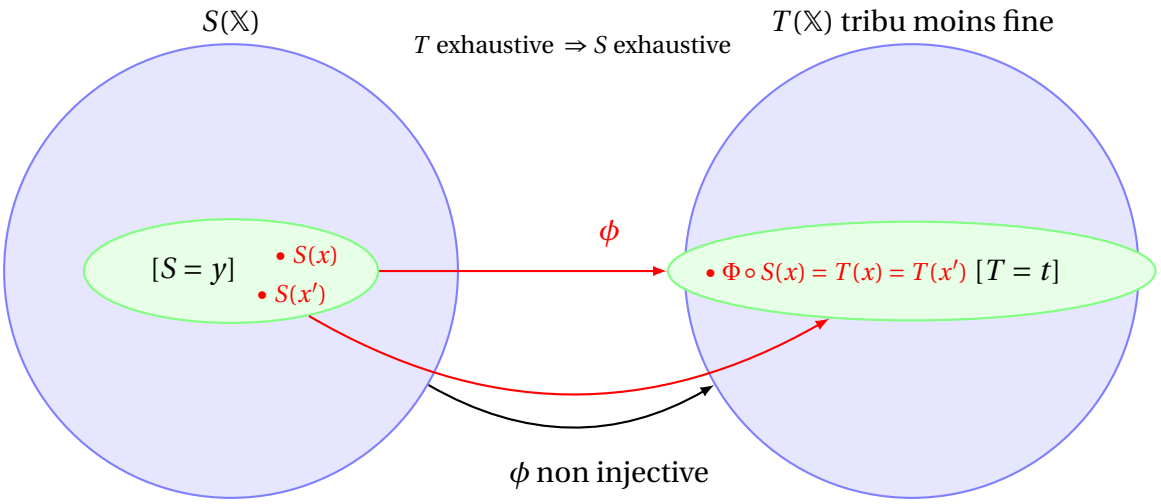

Figure 3.2 – Exhaustivité

Théorème 43 Critère de minimalité

Soit S une statistique d'un modèle dominé, telle que

$$LR(x,y,\theta) \text{ ne dépend pas de } \theta \iff S(x) = S(y) \tag{26}$$

Alors S est exhaustive et minimale.

Démonstration:

Soit S une statistique exhaustive; $L(x,\theta) = h(x)g(t,\theta)$ et si $S(x) = S(y)$,

$$LR(x,y,\theta) = \frac{h(x)}{h(y)} = \text{constante en } \theta \tag{27}$$

□

Exemple:

Reprenons à nouveau l'exemple du début de chapitre:

$$LR(x,y,\theta) = \theta^{S(x)-S(y)}(1-\theta)^{S(y)-S(x)} = \left(\frac{\theta}{1-\theta}\right)^{S(x)-S(y)} \tag{28}$$

qui est constant si, et seulement si, $S(x) = S(y)$. Donc $S(X) = \sum_{i=1}^{n} X_i$ est une statistique exhaustive minimale.

Exemple:

Reprenons l'exemple de l'échantillon uniforme sur $[\theta - 1/2, \theta + 1/2]$. En calculant le rapport de vraisemblance, on voit facilement (à faire) que le couple $(X_{(1)}, X_{(n)})$ est une statistique exhaustive minimale.

Théorème 44

Si le modèle est exponentiel, alors:

- La statistique canonique S est exhaustive.
- Si l'espace Λ des paramètres canoniques contient un ouvert de $\mathbb{R}^k$ ou un repère affine de $\mathbb{R}^k$, alors S est minimale

$$LR(x,y,\theta) = g(\theta) = \frac{h(x)}{h(y)} \exp\left(\langle \lambda(\theta), S(x) - S(y) \rangle\right) \tag{29}$$

dire que g est constante en θ est équivalent à dire que $\forall \lambda \in \Lambda, < \lambda, S(x) - S(y) >= c$, où c est une constante indépendante de θ, mais qui peut dépendre de x et y. Si Λ contient un ouvert, alors dire que ce produit scalaire

est constant sur tout l'ouvert est équivalent à dire que $< \lambda, S(x) - S(y) >= 0$ (en faisant la différence de deux produits scalaires pour des valeurs de λ et λ' différentes), ou encore que le vecteur $S(x) - S(y)$ est orthogonal à une infinité de vecteurs de l'espace vectoriel Λ, et l'on a donc $S(x) - S(y) = 0$. Même si Λ ne contient pas d'ouvert, le même argument est valable si l'on démontre que $< e_i, S(x) - S(y) >= 0$ pour une famille $(e_i)_i$ formant une base de Λ. Il suffit alors que le produit scalaire soit nul en chaque vecteur de la base pour conclure que le vecteur $S(x) - S(y) = 0$. Suppsons donc qu'il existe un repère affine $\lambda_0, ..., \lambda_n$ de Λ. Cela signifie qu'on peut trouver $n+1$ points non liés dans Λ; en ce cas, les vecteurs $e_i = \lambda_i - \lambda_0$, $i = 1, ..., n$ forment une base de Λ et si $\forall i = 0, ..., n$, $< \lambda_i, S(x) - S(y) >= c$ alors $< \lambda_i - \lambda_0, S(x) - S(y) >= 0$ et donc $S(x) = S(y)$. □

Exemple:

Soit $(X_1, ..., X_n)$ un n-échantillon de loi $\mathcal{N}(\theta, \theta^2)$. Il s'agit d'un modèle exponentiel avec pour paramètre $(1/\theta, -1/\theta^2)$ et pour statistique $(n\overline{x}, -n\overline{x^2}/2)$.

L'espace canonique Λ est formé par une courbe d'équation $u = -\lambda^2$ qui est fermée dans $\mathbb{R}^2$ et ne peut donc contenir aucun ouvert de $\mathbb{R}^2$. Par contre, on peut trouver trois points non alignés sur cette courbe qui forment un repère affine. On donc en déduire que la statistique est bien exhaustive minimale.

$$\Lambda = \left\{ \left(\frac{1}{\theta}, -\frac{1}{\theta^2} \right); \theta > 0 \right\} = \{(\lambda, -\lambda^2); \lambda > 0\} \subset \mathbb{R}^2 \tag{30}$$

On peut par contre montrer que si l'échantillon a pour loi $\mathcal{N}(\theta, \theta)$ (c'est à dire qu'on cherche à estimer la moyenne dans une loi où moyenne et écart type sont égaux), alors la statistique exhaustive précédente n'est pas minimale (en exercice).

3.1.2 Statistique libre et statistique complète

Une statistique exhaustive apporte toute l'information sur θ contenue dans X. Une statistique dont la loi ne dépend pas de θ n'apporte aucune information sur θ.

Définition 45

Une statistique T est libre vis à vis de θ si sa loi ne dépend pas de θ.
Elle est libre du premier ordre si la fonction (de θ) $\mathbb{E}_\theta[T]$ est constante.
Une statistique U est ignorable s'il existe une statistique exhaustive T stochastiquement indépendante de U. T apporte alors toute l'information sur θ et U apporte une information complémentaire à T. Si deux statistiques sont dépendantes, alors les informations qu'elles apportent sur un paramètre sont redondantes.

ignorable ⇒ libre, mais la réciproque est fausse

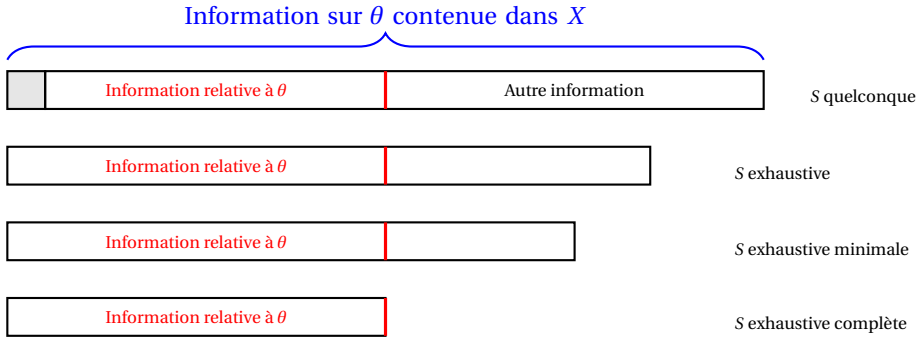

Figure 3.3 – Information, exhaustivité, minimalité, complétude.

Exemple:

Considérons un modèle d'échantillonnage gaussien $\mathcal{N}(\theta, 1)^{\otimes n}$. La statistique

$$S^2 = \sum_{i=1}^{n} \left(X_i - \overline{X} \right)^2 / n \tag{31}$$

est libre, car sa loi ne dépend pas de θ. En effet, en posant $Y_i = X_i - \theta$, $Y \sim \mathcal{N}(0, I_n)$ et

$$S^2 = \sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2 / n \tag{32}$$

Pour la même raison, la statistique d'étendue empirique $T = X_{(n)} - X_{(1)}$ est libre également.

Exemple:

Reprenons l'exemple de l'échantillon de loi uniforme sur $[\theta - 1/2, \theta + 1/2]$. On a vu que $T = (T_1, T_2) = (X_{(1)}, X_{(n)})$ est exhaustive minimale. Sa densité est donnée par

$$g(s,t) = n(n-1)(t-s)^{n-2} \mathbb{1}_{[\theta-1/2 \le s \le t \le \theta+1/2]} \tag{33}$$

$U = X_{(n)} - X_{(1)}$ est une statistique libre de θ car sa densité est

$$h(u) = n(n-1)u^{n-2} \mathbb{1}_{[0,1]}(u) \tag{34}$$

Pourtant, U n'est pas indépendante de T. T apporte donc une information inutile U sur θ.

Exemple:

Considérons un n-échantillon $X = (X_1, ..., X_n)$ dont la loi a pour densité

$$f(x) = \frac{1}{\ln \alpha} \frac{1}{x} \mathbb{1}_{]\theta, \alpha\theta[}(x) \tag{35}$$

pour $\alpha > 1$ et $\theta > 0$ inconnu. La statistique $(X_{(1)}, X_{(n)})$ est exhaustive minimale. Il en est de même de la statistique $T(X) = (X_{(n)} \div X_{(1)}, X_{(1)} X_{(n)})$, tandis que la statistique $U(X) = X_{(n)} \div X_{(1)}$ est libre et il est clair que T et U ne sont pas indépendantes.

Une statistique exhaustive minimale réduit au maximum les données sans perdre d'information sur θ. Mais une telle statistique n'est pas nécessairement indépendante de toute statistique libre, car une statistique libre peut compléter l'information apportée par une statistique minimale. La complétude assure que la statistique est bien indépendante de toute statistique libre: toute la liberté résiduelle a été extraite de la statistique et toute transformation la rend non libre. On cherche des statistiques exhaustives T ne contenant aucune composante libre, c'est à dire que pour toute fonction g non constante, $g(T)$ ne sera pas libre au premier ordre. Ou encore, Si l'espérance est constante, alors g est constante. Ces statistiques ne doivent contenir aucun matériel (c.-à-d. aucune fonction de T) libre, même du premier ordre. Cette notion est apportée par la complétude.

Définition 46

Une statistique T est complète s'il n'existe pas de fonction g non constante, intégrable, à valeurs dans $\mathbb{R}$ qui soit libre. Autrement dit,

$$\forall \theta,\ \mathbb{E}_\theta[g(T)] = 0 \Rightarrow g \equiv 0\ \mathbb{P}_\theta - \text{p.s.} \tag{36}$$

Théorème 47 Lehmann & Scheffé (1950), Bahadur (1957)

T exhaustive complète $\Rightarrow$ T exhaustive minimale.

Démonstration:

Soit T une statistique complète et S une statistique exhaustive minimale. Alors $S = g_1(T)$. Mais S est exhaustive; la loi conditionnelle de X sachant $[S = s]$ ne dépend pas de θ. En particulier, la loi conditionnelle de T sachant $[S = s]$ ne dépend pas de θ. Ainsi, $g_2(S) = \mathbb{E}_\theta[T|S]$ est fonction de S uniquement et ne dépend pas de θ. Soit $g(T) = T - g_2 \circ g_1(T)$. Pour tout θ, on a:

$$\mathbb{E}_\theta[g(T)] = \mathbb{E}_\theta[T] - \mathbb{E}_\theta[g_2(S)] = \mathbb{E}_\theta[T] - \mathbb{E}_\theta[\mathbb{E}_\theta[T|S]] = 0 \tag{37}$$

Mais T est complète, donc $g \equiv 0$ $\mathbb{P}_\theta$-p.s. pour tout θ. Donc $T = g_2(S)$ $\mathbb{P}_\theta$-p.s. pour tout θ. T est fonction d'une statistique exhaustive minimale, T est donc minimale.

□

Exemple:

On considère (encore) un n-échantillon de loi de Bernoulli de paramètre θ inconnu et l'on pose $T(X) = \sum_{i=1}^n X_i$ dont on a vu qu'elle est une statistique exhaustive minimale. Est-elle complète ? Considérons $g : \{0,...,n\} \longrightarrow \mathbb{R}$ telle que $\mathbb{E}_\theta[g(T)] = 0$. Comme la loi de T est une loi binomiale, on a

$$h(\theta) = \mathbb{E}_\theta[g(T)] = \sum_{t=0}^{n} g(t) \left(\begin{array}{c} n \\ t \end{array} \right) \theta^t (1-\theta)^{n-t} = 0 \tag{38}$$

h est donc un polynôme de degré n avec une infinité de zéros: il est nul et par suite T est complète.

Exemple:

On considère un n-échantillon de loi uniforme sur $[\theta - 1/2, \theta + 1/2]$ et $T = (X_{(1)}, X_{(n)})$. Alors T n'est pas complète (à faire en exercice).

Exemple:

Considérons un n-échantillon $(X_1,...,X_n)$ de loi de densité

$$f(x) = e^{x-\theta} \mathbb{1}_{]-\infty,\theta]}(x) \tag{39}$$

$X_{(n)} = \max_{i=1}^n X_i$ est une statistique exhaustive pour θ et sa densité est

$$h(u) = n e^{n(u-\theta)} \mathbb{1}_{]-\infty,\theta]}(u) \tag{40}$$

Soit g une fonction d'intégrale nulle par rapport à cette densité, pour tout θ. On a alors

$$\forall \theta \in \mathbb{R}, \int_{-\infty}^{\theta} g(u) n e^{n(u-\theta)} du = 0 \tag{41}$$

Soit encore

$$\forall \theta \in \mathbb{R}, \int_{-\infty}^{\theta} g(u) e^{nu} du = 0 \tag{42}$$

Ainsi, sur tout intervalle $[\theta, \theta']$,

$$\int_{\theta}^{\theta'} g(u) e^{nu} du = 0 \tag{43}$$

La fonction g est donc nulle presque partout (pourquoi ?).

Exemple: applications aux familles exponentielles.

On considère le modèle exponentiel dont la vraisemblance est donnée par

$$L(x,\theta) = h(x) \exp\big(\langle \lambda(\theta), T(x) \rangle - \beta(\theta)\big) \tag{44}$$

Avec $\Lambda = \{\lambda(\theta); \theta \in \Theta\}$ espace canonique des paramètres. Si Λ contient un ouvert non vide, alors S est exhaustive complète. En effet, pour une fonction g donnée,

$$\mathbb{E}_\theta[g(T)] = \int_{\mathbb{R}^k} g(t) e^{<\lambda,t>} d\mu(t) \tag{45}$$

Il s'agit de la transformée de Laplace de la fonction g. Les propriétés de la transformation de Laplace (dans $\mathbb{C}$) assurent alors que $g \equiv 0$ (principe des zéros isolés).

Exemple:

Dans l'exemple de l'échantillon de Bernoulli, la statistique canonique est $T(x) = \sum_i X_i$ et le paramètre canonique est $\lambda(\theta) = \ln(\theta/(1-\theta))$. Ainsi, $\Lambda = \{\lambda(\theta) = \ln(\theta/(1-\theta)), \theta \in]0,1[\}$. Une étude de fonction donne facilement $\Lambda(\Theta) = \mathbb{R}$ qui un ouvert de $\mathbb{R}$. On en déduit donc que T est complète.

Exemple:

Soit $(X_1,...,X_n)$ un n-échantillon de loi $\mathcal{N}(m,\sigma^2)$ avec $\theta = (m,\sigma^2)$. En posant $T_1 = \sum X_i$, $T_2 = -\sum X_i^2$, $T = (T_1,T_2)$, $\lambda = (m/\sigma^2, 1/(2\sigma^2))$, on voit que le modèle est celui d'une famille exponentielle de rang plein. Alors T est complète pour λ et également pour θ par bijectivité de $\lambda(\theta)$. Ainsi, $(\overline{X}, S^2)$ est une statistique complète pour $\theta = (m,\sigma^2)$.

Exemple et exercice:

Déterminer une statistique exhaustive minimale et complète pour le modèle multinomial.

Théorème 48 de Basu

Soit T une statistique exhaustive complète et U une statistique libre. Alors $T \perp\!\!\!\perp U$.

Démonstration:

Soit U une statistique libre et $\theta \in \Theta$ fixé. Pour tout ensemble mesurable $B \in \mathbb{X}$, $\mathbb{P}_\theta[U \in B]$ ne dépend pas de θ. Par exhaustivité de T, il existe également une version de la probabilité conditionnelle $\mathbb{P}[U \in B|T = t]$ sindépendante de θ. Or,

$$\mathbb{P}[U \in B] = \int \mathbb{P}[U \in B|T = y]P_T(dy) \tag{46}$$

En posant $f(y) = \mathbb{P}[U \in B] - \mathbb{P}[U \in B|T = y]$, on voit que l'intégrale de f est nulle. Mais f est bornée et T est complète. Donc $f \circ T \equiv 0$ $\mathbb{P}_\theta$-p.s. et les deux probabilités sont égales, ce qui assure que T et U sont indépendantes.

□

Exemple:

Le théorème de Basu donne un moyen rapide de démontrer que la moyenne et la variance empirique d'un échantillon gaussien sont indépendantes. En effet, $\overline{X}$ est exhaustive et complète pour la moyenne θ (c'est un modèle exponentiel). Par ailleurs S^2 est libre. D'après le théorème de Basu, les deux statistiques sont donc indépendantes.

3.2 Information de Fisher

3.2.1 Introduction et notations

L'information de Fisher tente de quantitier la quantité d'information sur un paramètre contenu dans un échantillon. Il existe plusieurs définitions différentes de la notion de quantité d'information. La plus ancienne et la plus connue est l'information au sens de Shannon (1948) qui permet de définir l'entropie d'une source aléatoire, le taux d'information d'un langage, etc. C'est le domaine de la théorie de l'information. L'information au sens de Kullback-Leibler en est une autre forme. L'information au sens de Fisher que nous allons voir maintenant n'est pas exactement le même objet, mais a des propriéts similaires.

L'information au sens de Fisher est une notion locale, qui évalue le pouvoir de discrimination du modèle entre deux valeurs proches de θ (c'est la courbure de la vraisemblance au voisinage de θ). Sa définition met donc en jeu des notions de convexité et de dérivée seconde par rapport à θ. Cela suppose donc que la vraisemblance soit définie et que sa différentielle seconde existe. Enfin, il faudra supposer que Θ est un ouvert de $\mathbb{R}^p$.

Soit $\theta = (\theta_1, ..., \theta_p) \in \Theta \subset \mathbb{R}^p$ ouvert.

Soit $g : \Theta \longrightarrow \mathbb{R}$ une fonction deux fois différentiable. On notera $\nabla g(\theta)$ son gradient en θ et $H_g(\theta)$ ou parfois $\nabla^2 g(\theta)$ sa matrice hessienne. On notera également pour un vecteur aléatoire colonne $S = (S_1, ..., S_p)' \in \mathbb{R}^p$, $\mathbb{V}(S)$ sa matrice de variance/covariance

$$\mathbb{V}(S) = \mathbb{E}\left[(S - \mathbb{E}[S]) \times (S - \mathbb{E}[S])'\right] = \begin{pmatrix} \text{cov}(S_1, S_1) & ... & \text{cov}(S_1, S_p) \\ \vdots & & \vdots \\ \text{cov}(S_p, S_1) & ... & \text{cov}(S_p, S_p) \end{pmatrix} \tag{47}$$

Nous noterons $l(X, \theta) = \ln L(X, \theta)$ la log-vraisemblance. Le score du modèle est (sous réserve d'existence),

$$S(X, \theta) = \frac{\partial}{\partial \theta} \ln L(X, \theta) = \nabla l(\theta) \tag{48}$$

C'est une fonction de X, donc une variable aléatoire. Par contre, comme S dépend de θ, ce n'est pas une statistique.

3.2.2 Notion de modèle régulier

Définition 49

Un modèle paramétrique est régulier si, et seulement si,

- Il est dominé, homogène et $\Theta \subset \mathbb{R}^p$ est un ouvert.
- $L(x,\theta) > 0 \ \forall x \in \mathbb{X}, \forall \theta \in \Theta$.
- $\theta \longrightarrow L(x,\theta)$ est de classe C^2 pour presque tout x.
- $\forall B \in \mathbb{X}, \theta \longrightarrow \int_B L(x|\theta) d\mu(x)$ est deux fois différentiable sous le signe somme.
- Le score $S(X,\theta) \in L^2(\mathbb{P}_\theta)$.

Les modèles gaussiens ou de Poisson sont réguliers, mais pas le modèle uniforme sur $[0,\theta]$.

Définition 50

L'information de Fisher $\mathbb{I}(\theta)$ du modèle est la variance du score:

$$\mathbb{I}(\theta) = \mathbb{V}_\theta(S) = \mathbb{V}_\theta(\nabla l(X,\theta)) \tag{49}$$

Propriété de l'information d'un modèle régulier

- $\mathbb{E}_\theta[S] = 0$
- $\mathbb{I}(\theta) = \mathbb{E}_\theta[SS^T] = -\mathbb{E}_\theta\left[\dfrac{\partial^2}{\partial\theta^2} l(X,\theta)\right] = -\mathbb{E}_\theta\left[H_l(X,\theta)\right]$

En dimension 1, $\mathbb{I}(\theta) = \mathbb{E}\left[S^2\right] = -\mathbb{E}\left[l''(\theta)\right]$.

Démonstration:

$$g(\theta) = \mathbb{E}_\mu[L(X,\theta)] = \int_{\mathbb{X}} L(x,\theta) d\mu(x) \tag{50}$$

Comme $L(.,\theta)$ est une densité de probabilité, $g(\theta)$ est une fonction constante en θ. D'après la seconde hypothèse, on peut dériver cette fonction en θ et l'on a donc $\nabla g(\theta) = 0$. Mais d'après l'hypothèse 3 (dérivation sous le signe somme), on a également

$$\nabla g(\theta) = \int_{\mathbb{X}} \frac{\nabla L(x,\theta)}{L(x,\theta)} L(x,\theta) d\mu(x) = \int_{\mathbb{X}} S(x,\theta) L(x,\theta) d\mu(x) = \mathbb{E}_\theta[S] = 0 \tag{51}$$

Le score est donc centré. En dérivant une seconde fois,

$$H_g(\theta) = \int_{\mathbb{X}} \left(H_l(x,\theta) + SS'\right) L(x,\theta) d\mu(x) = 0 \tag{52}$$

On en déduit donc la seconde propriété.

□

Exemple:

Considérons un n-échantillon $X = (X_1, ..., X_n)$ de loi de Benoulli de paramètre θ et considérons la statistique $T(X) = \sum_{i=1}^n X_i$. La vraisemblance du modèle est

$$L(X,\theta) = \exp\left(T(X)\ln\theta + (n - T(X)\ln(1-\theta)\right) \tag{53}$$

Considérons une observation $x \in \mathbb{R}^n$; x est un vecteur de dimension n et $\theta \in]0,1[$ est un paramètre scalaire. Le score du modèle est

$$S(x,\theta) = \frac{\partial l(x,\theta)}{\partial\theta} = \frac{T(x)}{\theta} - \frac{n - T(x)}{1-\theta} = \frac{T(x)}{\theta(1-\theta)} - \frac{n}{1-\theta} \tag{54}$$

On en déduit l'information du modèle

$$\mathbb{I}(\theta) = \mathbb{V}\left(S(X,\theta)\right) = \frac{n}{\theta(1-\theta)} \tag{55}$$

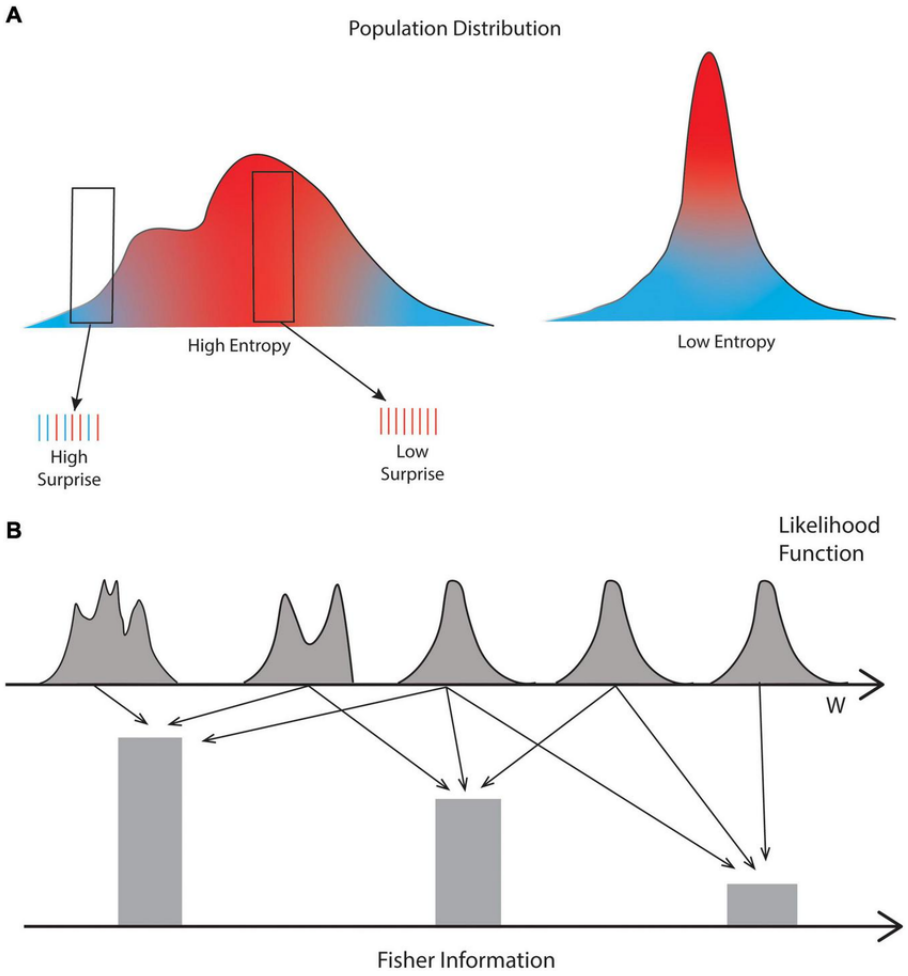

Figure 3.4 – Information de Fisher et de Shannon. A: l'entropie de Shannon capture l'étalement (en rouge et bleu) de la statistique en terme de vraisemblance. B: l'information de Fisher capture la quantité d'information que la vraisemblance apporte sur la vraie valeur du paramètre, grâce à la variation de la fonction de vraisemblance. $\mathbb{I}(\theta) = -\mathbb{E}\left[l''(\theta)\right]$ (R. Grzywacz, H. Aleem).

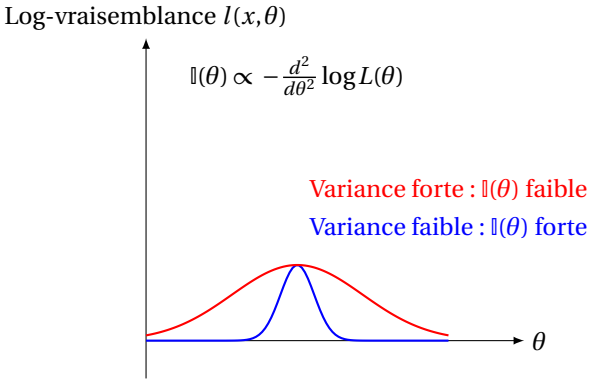

Figure 3.5 – Information de Fisher : $\mathbb{I}(\theta) = -\mathbb{E}\left[l''(\theta)\right]$. En rouge : grande variance donc peu d'information en θ, en bleu : variance faible, donc forte information sur θ au voisinage du maximum de $l(\theta)$.

3.2.3 Propriétés de l'information de Fisher.

Information de Fisher apportée par une statistique

Soient $\left(\mathbb{X},\mathfrak{F},(\mathbb{P}_\theta)_{\theta\in\Theta}\right)$ un modèle régulier et T une statistique pour laquelle le modèle image est également régulier. Soit $g(t,\theta)$ la densité de T. L'information de Fisher apportée par T est (par définition) l'information de Fisher du modèle image:

$$I_T(\theta) = \mathbb{V}\left(S(T,\theta)\right) \tag{56}$$

Le score du modèle image par T est

$$\frac{d}{d\theta}\ln g(T,\theta) \tag{57}$$

Ce résultat est une conséquence directe du théorème de transfert.

Additivité de l'information

Si T_1 et T_2 sont deux statistiques indépendantes alors

$$I_{(T_1,T_2)}(\theta) = I_{T_1}(\theta) + I_{T_2}(\theta) \tag{58}$$

En particulier, dans un modèle d'échantillonnage régulier $X = (X_1,...,X_n)$,

$$\mathbb{I}_X(\theta) = n \times \mathbb{I}_{X_1}(\theta) \tag{59}$$

Ce résultat est fondamental et la notion de quantité d'information a été construite sur la base de cette propriété: on souhaite que l'information soit une quantité additive. Deux phénomènes aléatoires indépendants doivent apporter une quantité d'information totale égale à la somme de leur quantité d'information respective. C'est une conséquence directe du fait que le logarithme transforme un produit en somme.

Soit T une statistique de densité $g(t,\theta)$ et $h(x,t,\theta)$ la densité conditionnelle de X sachant $[T=t]$. Alors

$$I(\theta) = I_T(\theta) + I_{X|T}(\theta) \tag{60}$$

où $I_{X|T}$ est l'information de Fisher de X sachant T.

Cette propriété illustre également la dégradation de la quantité d'information: $\mathbb{I}(\theta) \geq \mathbb{I}_T(\theta)$. Si $\theta\in\mathbb{R}$, $\mathbb{I}_{X|T}(\theta) > 0$. Si $\theta\in\mathbb{R}^p$, $\mathbb{I}_{X|T}(\theta)\in \mathrm{Sym}^+(\mathbb{R}^p)$.

On rappelle que pour des matrices, $A\geq B \iff \forall\theta\in\mathbb{R}^p,\ \theta' A\theta \geq \theta' B\theta$.

Enfin, si T est une statistique exhaustive de densité $g(t,\theta)$ et $h(x,t)$ est la densité conditionnelle est X sachant $[T=t]$ (elle ne dépend donc pas de θ) alors

$$I_X(\theta) = I_T(\theta) \tag{61}$$

C'est même une équivalence, sous hypothèse de régularité: T exhaustive $\iff I(\theta) = I_T(\theta)$. Pour une statistique libre de θ, la quantité d'information apportée est nulle (c'est également une équivalence).

Exemple:

Considérons un n-échantillon gaussien $(X_1,...,X_n)$ d'une v.a. $X\sim\mathcal{N}(m,\sigma^2)$. Le paramètre $\theta=(m,\sigma^2)$ est vectoriel. Le modèle est régulier et la log-vraisemblance d'une observation x est

$$l(x,\theta) = -\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}(x-m)^2 \tag{62}$$

On a alors (poser $v=\sigma^2$ pour simplifier les calculs) :

$$\frac{\partial^2 l}{\partial m^2}(x,\theta) = -\frac{1}{\sigma^2} \quad \frac{\partial^2 l}{\partial(\sigma^2)^2}(x,\theta) = \frac{1}{2\sigma^4} - \frac{(x-m)^2}{\sigma^6} \quad \frac{\partial^2 l}{\partial m\partial(\sigma^2)}(x,\theta) = -\frac{x-m}{\sigma^4} \tag{63}$$

Ainsi,

$$\mathbb{I}_X(\theta) = -\mathbb{E}\left[\begin{pmatrix} \frac{\partial^2 l}{\partial m^2}(X,\theta) & \frac{\partial^2 l}{\partial m\partial(\sigma^2)}(X,\theta) \\ \frac{\partial^2 l}{\partial m\partial(\sigma^2)}(X,\theta) & \frac{\partial^2 l}{\partial(\sigma^2)^2}(X,\theta) \end{pmatrix}\right] = \begin{pmatrix} 1/\sigma^2 & 0 \\ 0 & 1/(2\sigma^4) \end{pmatrix} \tag{64}$$

L'information de Fisher associée au n-échantillon sera donc

$$\mathbb{I}(\theta) = \begin{pmatrix} n/\sigma^2 & 0 \\ 0 & n/(2\sigma^4) \end{pmatrix} \tag{65}$$

reparamétrisation

Considérons un nouveau paramètre $\lambda = \phi(\theta)$. On peut calculer l'information de Fisher du modèle relativement à ce nouveau paramètre via la formule suivante :

$$\mathbb{I}_X(\lambda) = J' \times \mathbb{I}_X(\theta) \times J = \left(\nabla \phi^{-1}(\lambda)\right)' \times \mathbb{I}_X(\theta) \times \left(\nabla \phi^{-1}(\lambda)\right) \tag{66}$$

où J est la matrice jacobienne de l'application inverse de ϕ. Pour exprimer cette nouvelle information en fonction de λ, il faut pouvoir exprimer θ en fonction de λ en inversant la fonction ϕ.

En dimension 1, si θ et λ sont des paramètres réels, la formule devient:

$$\mathbb{I}_X(\lambda) = \frac{\mathbb{I}_X(\theta)}{\phi'(\theta)^2} \tag{67}$$

Exemple:

Partant d'un échantillon de Poisson de paramètre λ, on s'intéresse au paramètre $\theta = \exp(-\lambda)$. Déterminer l'information de Fisher de l'échantillon relativement à ce nouveau paramètre.

3.2.4 Information de Kullback

Définition 52

$$K(\theta_0, \theta_1) = \mathbb{E}_{\theta_0}\left[\ln \frac{L(X,\theta_0)}{L(X,\theta_1)}\right] = \int_{\mathbb{R}^n} \ln \frac{L(x,\theta_0)}{L(x,\theta_1)} L(x,\theta_0)\, dx \tag{68}$$

$\ln(L(X,\theta_0)/L(X,\theta_1))$ est le pouvoir discriminant de θ_0 contre θ_1.

Propriété

$$\left.\frac{\partial^2 K(\theta_0,\theta)}{\partial \theta^2}\right|_{\theta=\theta_0} = \mathbb{I}(\theta_0) \tag{69}$$

Exemple:

Si X suit une loi de Poisson $\mathscr{P}(\theta)$, alors $K(\theta_0,\theta_1) = \theta_1 - \theta_0 + \theta_0 \ln(\theta_0/\theta_1)$.

3.3 Estimateurs optimaux

3.3.1 Fonction de coût et risque moyen

Une fonction de coût mesure l'écart (la perte) entre deux quantités. Typiquement, la qualité d'un estimateur va être appréciée en fonction de l'écart qui existe entre l'estimateur T et le paramètre θ à estimer. Une fonction de coût c doit êtes positive et vérifier $\theta = \theta' \Rightarrow c(\theta,\theta') = 0$.

Toute distance sur Θ vérifie en particulier ces conditions. Les trois fonctions de coût les plus couramment utilisées sont le coût quadratique c_2, le coût absolu c_1 et le coût 0-1 noté c_0:

$$c_2(\theta,\theta') = (\theta-\theta')^2 \qquad c_1(\theta,\theta') = |\theta-\theta'| \qquad c_0(\theta,\theta') = \mathbb{1}_{[|\theta-\theta'|>\epsilon|]} \tag{70}$$

Définition 54

Le risque moyen d'un estimateur T de θ est l'espérance de la fonction de coût entre T et θ:

$$R(T,\theta) = \mathbb{E}_\theta[c(T,\theta)] \tag{71}$$

Un estimateur est préférable à un autre (à θ fixé) si son risque moyen est inférieur. Il est uniformément préférable si le risque est inférieur quelque soit $\theta \in \Theta$.

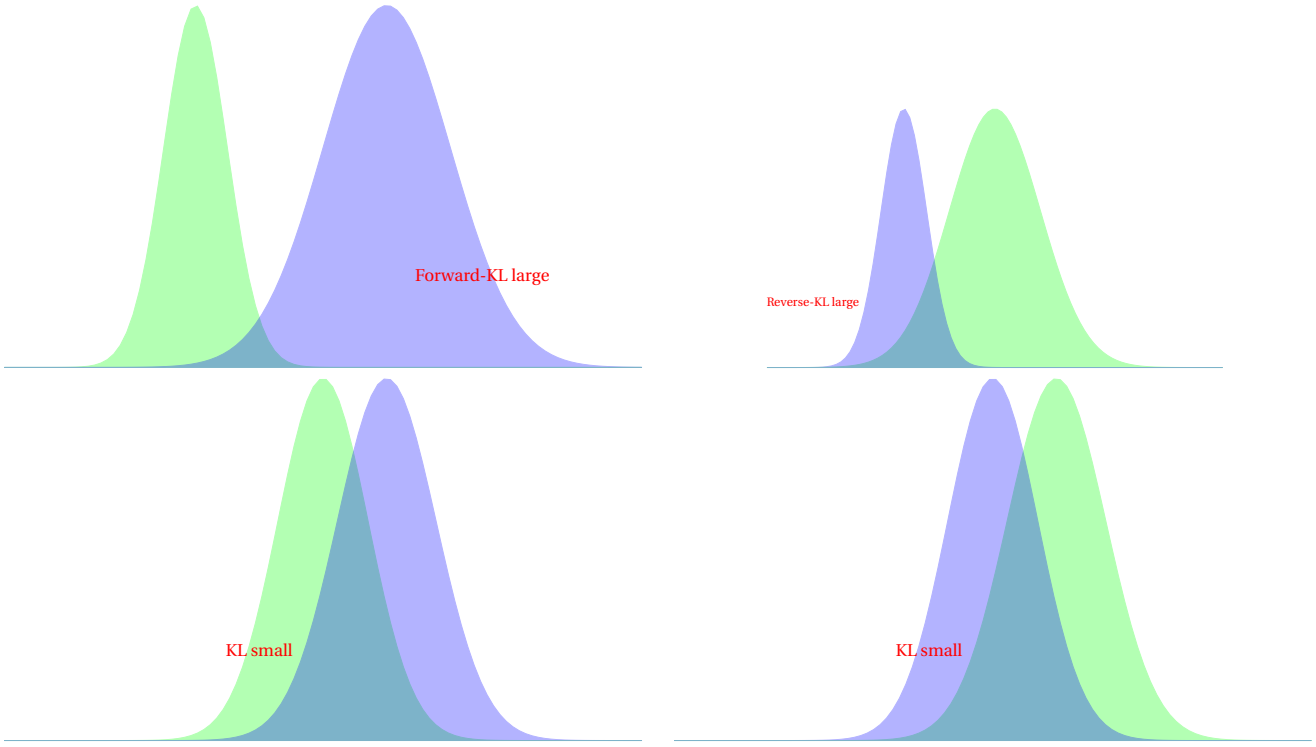

Figure 3.6 – Information de Kullabck-Leibler.

Définition 55

Étant donné une classe $\mathcal{T}$ d'estimateurs de θ, $T \in \mathcal{T}$ est admissible dans $\mathcal{T}$ pour θ s'il est uniformément préférable à tout autre estimateur de $\mathcal{T}$:

$$\forall T' \in \mathcal{T}, \forall \theta \in \Theta,\ R(T,\theta) \leq R(T',\theta) \tag{72}$$

3.3.2 Estimateur VUMSB

On souhaite déterminer l'estimateur admissible pour la classe des estimateurs sans biais. La formule reliant le biais et la variance nous donne

$$R(T,\theta) = \mathbb{V}_\theta(T) + (\mathbb{E}_\theta[T] - \theta)^2 \tag{73}$$

Définition 56

Un estimateur T^* est de variance uniformément minimale parmi les estimateurs sans biais de θ (on dira VUMSB) si T^* est admissible pour le risque quadratique, dans la classe des estimateurs sans biais. C'est à dire:

$$\bullet\ \mathbb{E}[T^*] = \theta \tag{74}$$

$$\bullet\ \forall\ T \text{ sans biais }, \mathbb{V}_\theta(T^*) \leq \mathbb{V}_\theta(T) \tag{75}$$

Lorsqu'il existe, l'estimateur VUMSB est unique $\mathbb{P}_\theta$-p.s. pour tout θ. En effet, si deux estimateurs sont VUMSB, leur moyenne est égale à θ et leur variance commune. La moyenne arithmétique de ces deux estimateurs est également sans biais, avec une variance strictement plus faible, ce qui est absurde, sauf si les deux estimateurs sont p.s. égaux.

Théorème 57 de Rao-Blackwell

Soit T un estimateur sans biais et S une statistique exhaustive. Alors

$$T^* = \mathbb{E}_\theta[T|S] \tag{76}$$

est un estimateur sans biais de θ préférable à T pour le risque moyen quadratique.

Démonstration:

Comme S est exhaustive, T^* est bien une statistique car la loi conditionnelle de X sachant $[S=s]$ ne dépend pas de θ. Soit $h(x,s)$ cette densité.

$$T^*(s) = \mathbb{E}_\theta[T|S=s] = \int T(x)h(x,s)d\mu(s) \tag{77}$$

T^* est sans biais car $\mathbb{E}_\theta[T^*] = \mathbb{E}_\theta[\mathbb{E}_\theta[T|S]] = \mathbb{E}_\theta[T] = \theta$.

$$\mathbb{V}_\theta(T) = \mathbb{E}_\theta\left[(T-T^*+T^*-\mathbb{E}_\theta[T])^2\right] = \mathbb{E}_\theta\left[(T-T^*)^2\right] + 2\mathbb{E}_\theta\left[(T-T^*)(T^*-\theta)\right] + \mathbb{E}_\theta\left[(T^*-\theta)^2\right] \tag{78}$$

$$= \mathbb{V}_\theta(T^*) + \mathbb{E}_\theta\left[(T-T^*)^2\right] \geq \mathbb{V}_\theta(T^*) \tag{79}$$

En effet,

$$\mathbb{E}_\theta\left[(T-T^*)(T^*-\theta)\right] = \mathbb{E}_\theta\left[\mathbb{E}_\theta\left[(T-T^*)(T^*-\theta)|S\right]\right] \tag{80}$$

$$= \mathbb{E}_\theta\left[(T^*-\theta)\,\mathbb{E}_\theta\left[(T-T^*)|S\right]\right] = 0 \tag{81}$$

□

Lorsque T est VUMSB et que S est exhaustive, alors T^* est uniformément préférable à T et T^* est aussi VUMSB. On a donc $T = T^*$.

Théorème 58 de Lehmann-Scheffé

Soit T un estimateur sans biais et S une statistique complète. Alors $T^* = \mathbb{E}_\theta[T|S]$ est VUMSB.

Démonstration:

Soient T_1 et T_2 deux estimateurs sans biais de θ et T_1^*, T_2^* les estimateurs optimaux correspondants. $T_1^* - T_2^*$ est un estimateur de moyenne nulle. Mais comme S est complète, on a alors $T_1^* = T_2^*$ $\mathbb{P}_\theta$-p.s. et on donc unicité.

Le fait que T^* soit VUMSB est une conséquence du théorème de Rao-Blackwell: supposons que T soit VUMSB. Alors T^* est préférable à T donc T^* est également VUMSB. Comme il est unique, on a égalité entre T et T^*. □

Exemple:

Pour un n-échantillon $(X_1,...,X_n)$ gaussien de loi $\mathcal{N}(m,\sigma^2)$, avec $\theta = (m,\sigma^2)$, la vraisemblance s'écrit

$$L(x,\theta) = (2\pi)^{-n/2}\exp\left[-\frac{n}{2\sigma^2}\overline{x^2} + \frac{nm}{\sigma^2}\overline{x} - \frac{nm^2}{2\sigma^2} - \frac{n}{2}\ln\sigma^2\right] \tag{82}$$

Le modèle est exponentiel (donc homogène) et la statistique naturelle est $\left(\overline{X^2},\overline{X}\right)$ et le paramètre canonique est $(-n/(2\sigma^2), nm/\sigma^2)$. L'espace des paramètres canoniques est $\Lambda =]-\infty,0[\times\mathbb{R}$. C'est donc un ouvert et par conséquent la statistique canonique est exhaustive et complète.

Posons $T = \left(\overline{X},\overline{X^2}\right)$. On peut donc choisir comme estimateur sans biais $\left(\overline{X},\widetilde{S}^2\right)$ avec $\widetilde{S}^2 = n/(n-1)\times\left(\overline{X^2}-\overline{X}^2\right)$. Le théorème de Lehmann-Scheffé donne alors comme VUMSB

$$\left(\mathbb{E}_\theta\left[\overline{X}|T\right], \mathbb{E}_\theta\left[\widetilde{S}^2|T\right]\right) \tag{83}$$

Mais $\overline{X}$ et $\widetilde{S}^2$ sont déjà des fonctions de T et par suite, $\left(\overline{X},\widetilde{S}^2\right)$ est un VUMSB pour θ.

3.3.3 Borne de Fréchet, Darmois, Cramer, Rao

Lorsque l'on utilise le risque quadratique multidimensionnel pour la classe des estimateurs sans biais, on peut obtenir une borne inférieure pour la variance.

Théorème 59 — Borne FDCR

On considère un modèle régulier $X \sim \mathbb{P}_\theta$ où l'information de fisher $\mathbb{I}(\theta)$ est inversible $\forall \theta \in \Theta$.
Soit T un estimtateur sans biais de $\phi(\theta) \in \mathbb{R}^d$ tel que :

- $\phi(\theta) = \mathbb{E}_\theta[T]$ est différentiable en θ.
- $\mathbb{E}_\theta[T]$ est différentiable en θ sous le signe somme.

Alors

$$\mathbb{V}_\theta(T) \geq \nabla\phi(\theta) \times \mathbb{I}(\theta)^{-1} \times \nabla\phi(\theta)' \tag{84}$$

Dans le cas particulier où T est un estimateur sans biais de θ, on a $\mathbb{E}_\theta[T] = \theta$, c'est à dire que la fonction ϕ se réduit à l'identité. L'inégalité devient alors

$$\mathbb{V}_\theta(T) \geq \mathbb{I}(\theta)^{-1} \tag{85}$$

Démonstration:

- Nous commençons la démonstration dans le cas de la dimension 1 ($d = p = 1, \theta \in \mathbb{R}, \phi(\theta) \in \mathbb{R}$). En dérivant par rapport à θ la fonction $g(\theta)$, on a:

$$\nabla\phi(\theta) = \int T(x)\nabla L(x,\theta)\frac{1}{L(x,\theta)}L(x,\theta)\mu(dx) = \mathbb{E}_\theta[T(X)S(X,\theta)] \tag{86}$$

où μ est la mesure dominante. Mais $\mathbb{E}_\theta[S] = 0$ donc:

$$\nabla\phi(\theta) = \mathbb{E}_\theta[TS] - \mathbb{E}_\theta[T]\mathbb{E}_\theta[S] = \mathrm{cov}_\theta(T,S) \tag{87}$$

Par l'inégalité de Cauchy-Schwarz,

$$\phi'(\theta)^2 = \mathrm{cov}_\theta(T,S)^2 \leq \mathbb{V}_\theta(T)\mathbb{V}_\theta(S) \tag{88}$$

donc

$$\phi'(\theta)^2\mathbb{I}(\theta)^{-1} \leq \mathbb{V}_\theta(T) \tag{89}$$

ce qui prouve l'inégalité.

- Soit maintenant $p > 1$. Pour simplifier les notations, on va omettre la variable θ, bien que toutes les expressions en dépendent. Posons $Z = T - \phi - \nabla\phi \times \mathbb{I}^{-1} \times S$. Z est une v.a. dont la matrice de variance est symétrique et positive. Nous allons démontrer que cette variance vérifie

$$\mathbb{V}(Z) = \mathbb{V}(T) - \nabla\phi \times \mathbb{I}^{-1} \times \nabla\phi' \tag{90}$$

de sorte que, comme $\mathbb{V}(Z) \geq 0$, l'inégalité sera démontrée. On rappelle que $\mathbb{I}$, ϕ et $\nabla\phi$ sont déterministes et que seuls S et T sont aléatoires. Par définition de la variance, on a

$$\begin{aligned}
\mathbb{V}(Z) &= \mathbb{E}\left[\left(T - \phi - \nabla\phi\,\mathbb{I}^{-1}S\right) \times \left(T - \phi - \nabla\phi\,\mathbb{I}^{-1}S\right)'\right] && (91)\\
&= \mathbb{E}\left[TT'\right] + \mathbb{E}\left[\nabla\phi\,\mathbb{I}^{-1}SS'\mathbb{I}^{-1}\nabla\phi'\right] - \mathbb{E}\left[(T-\phi)S'\,\mathbb{I}^{-1}\nabla\phi'\right] - \mathbb{E}\left[\nabla\phi\,\mathbb{I}^{-1}S(T-\phi)'\right] && (92)\\
&= \mathbb{V}(T) + \nabla\phi\,\mathbb{I}^{-1} \times \mathbb{E}[SS'] \times \mathbb{I}^{-1}\nabla\phi' - \mathbb{E}\left[(T-\phi)S'\right]\mathbb{I}^{-1}\nabla\phi' - \nabla\phi\,\mathbb{I}^{-1}\mathbb{E}\left[S(T-\phi)'\right] && (93)\\
&= \mathbb{V}(T) + \nabla\phi\,\mathbb{I}^{-1}\nabla\phi' - \left(\mathbb{E}[TS'] - \phi\mathbb{E}[S']\right)\mathbb{I}^{-1}\nabla\phi' - \nabla\phi\,\mathbb{I}^{-1}\left(\mathbb{E}\left[ST'\right] - \mathbb{E}[S]\phi'\right) && (94)\\
&= \mathbb{V}(T) - \nabla\phi\,\mathbb{I}^{-1}\nabla\phi' && (95)
\end{aligned}$$

□

Définition 60

Un estimateur T de $\phi(\theta)$ est efficace s'il est sans biais et atteint la borne FDCR.

Exemple:

Reprenons l'échantillon gaussien de l'exemple précédent, pour un paramètre $\theta = (m, \sigma^2)$. La borne de Cramer-Rao est donnée par

$$\mathbb{I}(\theta)^{-1} = \begin{pmatrix} \sigma^2/n & 0 \\ 0 & 2\sigma^4/n \end{pmatrix} \tag{96}$$

Par ailleurs, nous avons déjà calculé la matrice de covariance de l'estimateur $T = (\overline{X}, \widetilde{S}^2)$ qui vaut

$$\mathbb{V}_\theta(T) = \begin{pmatrix} \sigma^2/n & 0 \\ 0 & 2\sigma^4/(n-1) \end{pmatrix} \tag{97}$$

(le calcul de la variance de la variance empirique est assez lourd...). L'estimateur du maximum de vraisemblance n'est donc pas efficace pour θ. Par contre $\overline{X}$ est efficace pour la moyenne $m = \phi(\theta)$ où ϕ est la projection sur la première coordonnée. S^2 n'est pas efficace pour $\sigma^2 = \psi(\theta)$ projection sur la seconde coordonnée.

Théorème 61 — Efficacité asymptotique de l'e.m.v.

Soit $X_1, ..., X_n$ un n-échantillon de $X \sim \mathbb{P}_\theta$, $\theta \in \Theta$, où $f(x,\theta)$ est la densité de $\mathbb{P}_\theta$ par rapport à μ. On suppose que:

- Le modèle est homogène et identifiable.
- $\ln f(x,\theta)$ est de classe C^2 en θ et sa différentielle seconde est localement holdérienne:

$$\forall \theta \in \Theta, \exists\, V \text{ voisinage de } \theta \text{ et } M : \mathbb{X} \to \mathbb{R}_+ : \tag{98}$$

$$\forall \theta_1, \theta_2 \in V : \left|\left| \frac{d^2}{d\theta^2} \ln f(x,\theta_1) - \frac{d^2}{d\theta^2} \ln f(x,\theta_2) \right|\right| \leq M(x)|| \theta_1 - \theta_2||, \ \mathbb{P}_\theta \text{p.s.} \tag{99}$$

- Θ est un ouvert de $\mathbb{R}^p$ et $\mathbb{I}(\theta)$ est inversible $\forall \theta \in \Theta$

ALORS

L'estimateur $\widehat{\theta}_n$ du maximum de vraisemblance de θ est $\sqrt{n}$-consistant et asymptotiquement normal:

$$\sqrt{n}\left(\widehat{\theta}_n - \theta\right) \overset{\mathbb{P}_\theta}{\rightsquigarrow} \mathcal{N}(0, \mathbb{I}(\theta)^{-1}) \tag{100}$$

En particulier $\widehat{\theta}_n$ est asymptotiquement sans biais et efficace.

Démonstration:

On notera $\widehat{\theta}_n = \widehat{\theta}$ pour alléger les notations. Par définition de l'e.m.v.,

$$\nabla l(X, \widehat{\theta}) = \frac{\partial}{\partial \theta} l(X, \widehat{\theta}) = 0 \tag{101}$$

En effectuant le développement de Taylor à l'ordre 2 de cette expression, il existe $\zeta \in [\theta, \widehat{\theta}]$ tel que

$$0 = \frac{\partial}{\partial \theta} l(X, \widehat{\theta}) + (\widehat{\theta} - \theta) \frac{\partial^2}{\partial \theta^2} l(X, \zeta) \tag{102}$$

Notons

$$H_n = H_n(X,\theta) = \frac{1}{n} \frac{\partial^2}{\partial \theta^2} l(X,\zeta) \text{ et } O_n = O_n(X,\theta) = \frac{1}{\sqrt{n}} \frac{\partial}{\partial \theta} l(X, \widehat{\theta}) \tag{103}$$

On a alors

$$0 = O_n + \sqrt{n}(\widehat{\theta} - \theta) H_n \Rightarrow \sqrt{n}(\widehat{\theta} - \theta) = -H_n^{-1} O_n \tag{104}$$

- Convergence de O_n:

$$O_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{\partial}{\partial \theta} \ln f(X_i, \theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^n S_i \tag{105}$$

où S_i est le score du modèle image par X_i. Les S_i sont donc centrés et i.i.d. de variance $\mathbb{I}(\theta)$. D'après le théorème de la limite centrale, O_n converge en loi vers $\mathcal{N}(0, \mathbb{I}(\theta))$.

• Convergence de H_n:

$\widehat{\theta}$ est fortement consistant. Pour n assez grand, il existe un voisinage V de ζ inclus dans $[\theta,\widehat{\theta}]$ et vérifiant la condition 2:

$$\left|\left|H_n - \frac{d^2}{d\theta^2}\ln L(X,\theta)\right|\right| \leq ||\,\widehat{\theta}-\theta|| \times \frac{1}{n}\sum_{i=1}^{n} M(X_i) \tag{106}$$

Le premier terme à droite de l'inégalité est un $o_{\mathbb{P}}(1)$ et le second est un $\mathcal{O}_{\mathbb{P}}(1)$. Le produit des deux est donc un $o_{\mathbb{P}}(1)$. Ainsi, H_n a le même comportement asymptotique

$$\frac{1}{n}\frac{d^2}{d\theta^2}\ln L(X,\theta) = \frac{1}{n}\sum_{i=1}^{n}\frac{d^2}{d\theta^2}\ln f(X_i,\theta) = \frac{1}{n}\sum_{i=1}^{n} W_i \tag{107}$$

où les W_i sont intégrables et vérifient $\mathbb{E}_\theta[W_i] = \mathbb{I}(\theta)$. Les W_i sont i.i.d. car les X_i le sont et la loi forte des grands nombres assure la convergence presque sûre de H_n vers $-\mathbb{I}(\theta)$.

• Conclusion:

$\mathbb{I}(\theta)$ est inversible. En utilisant les deux résultats précédents et le lemme de Slutsky, on conclut que

$$\sqrt{n}\left(\widehat{\theta}_n - \theta\right) = -H_n^{-1}O_n \overset{\mathbb{P}_\theta}{\rightsquigarrow} \mathcal{N}(0,\mathbb{I}(\theta)^{-1}) \tag{108}$$

□

Pour terminer ce paragraphe sur l'estimation sans biais, une remarque:

Théorème 62 — Second principe fondamental de la statistique

Mieux vaut un petit biais qu'une grosse variance.

3.3.4 Estimateurs bayésiens

Dans un cadre bayésien, la comparaison des deux estimateurs se fera en intégrant le risque sur l'espace des paramètres. On parle de « risque intégré »

Définition 63

Le risque moyen *a posteriori* de l'estimateur T associé à la loi *a priori* Π et à l'observation x est

$$R(\Pi,T,x) = \mathbb{E}[c(H,T(x))|X=x] = \int_\Theta c(\theta,T(x))L(\theta,x)\,d\theta \tag{109}$$

Contrairement au cas fréquentiste, ce risque bayésien est un ordre total sur l'ensemble de tous les estimateurs.

Définition 64

Le risque intégré de l'estimateur T associé à la loi *a priori* Π est

$$R(\Pi,T) = \mathbb{E}[c(H,T)] = \int_{\mathbb{X}}\int_\Theta c(\theta,T(x))L(\theta,x)\,d\theta\,d\mu(x) \tag{110}$$

Si $\widehat{\theta}(x)$ est un estimateur minimisant le risque moyen *a posteriori*, alors $\widehat{\theta}$ minimise le risque intégré.

Définition 65

Soit $\mathcal{T}$ une classe d'estimateurs de θ. Le risque de Bayes est:

$$R(\Pi) = \inf_{T \in \mathcal{T}} R(\Pi, T) \tag{111}$$

L'estimateur de Bayes $\widehat{\theta}_\Pi$ associé à la loi *a priori* Π est l'estimateur qui minimise le risque intégré. Il vérifie:

- $R(\Pi, \widehat{\theta}) = R(\Pi)$ (112)
- $\widehat{\theta} = \text{argmin}_T R(\Pi, T)$ (113)

Pour une fonction de coût quadratique, l'estimateur bayésien est la moyenne *a posteriori*:

$$\widehat{\theta}(x) = \mathbb{E}[H|X = x] \tag{114}$$

En effet, soit $t = T(x)$. On veut minimiser en t:

$$g(t) = R(\Pi, T|x) = \mathbb{E}[(H-t)^2|X = x] = \mathbb{E}[H^2|X = x] - 2t\mathbb{E}[H|X = x] + t^2 \tag{115}$$

La résolution de $g'(t) = 0$ donne le résultat.

Pour le risque absolu, l'estimateur bayésien est la médiane *a posteriori*:

$$\widehat{\theta}(x) \text{ vérifie } \mathbb{P}[H \leq \widehat{\theta}(x)|X = x] = 1/2 \tag{116}$$

En effet:

$$g(t) = R(\Pi, T|x) = \mathbb{E}[(H-t)|X = x] = \int_{\mathbb{R}} |\theta - t| L(\theta, x) d\theta \tag{117}$$

$$= \int_t^\infty (\theta - t) L(\theta, x) d\theta + \int_{-\infty}^t (t - \theta) L(\theta, x) d\theta \tag{118}$$

$$= \int_t^\infty \mathbb{P}[H > \theta|X = x] d\theta + \int_{-\infty}^t \mathbb{P}[H < \theta|X = x] d\theta \tag{119}$$

$g'(t) = 0 \iff \mathbb{P}[H > t|X = x] = \mathbb{P}[H \leq t|X = x] = 1/2$

Chapitre 4

Estimation par intervalle de confiance

4.1 Intervalle et région de confiance: définitions

Définition 66

Soient T_1, T_2 deux statistiques à valeurs réelles telles que $T_1 \leq T_2$ $\mathbb{P}_\theta$-p.s. pour tout $\theta \in \Theta \subset \mathbb{R}$ et

$$\mathbb{P}_\theta[T_1 \leq \theta \leq T_2] = 1 - \alpha \tag{1}$$

Alors $IC_{1-\alpha}(\theta) = [T_1, T_2]$ est un intervalle de confiance de niveau $1-\alpha$ de θ.

Définition 67

Un sous-ensemble aléatoire $C(X)$ de $\Theta \subset \mathbb{R}^p$ avec $p \geq 2$, vérifiant pour tout θ:

$$\mathbb{P}_\theta[\theta \in C(X)] = 1 - \alpha \tag{2}$$

est appelée région de confiance de θ au niveau $1-\alpha$.

Les deux critères de qualité d'un intervalle de confiance (sa longueur et son niveau de confiance) s'opposent et il faut donc réaliser un compromis entre les deux. On cherche généralement un intervalle de longueur la plus petite possible, pour un niveau de risque α donné.

Étant donné la fonction de répartition F d'une v.a. X sur $\mathbb{R}$, on rappelle que le quantile d'ordre $\alpha \in]0,1[$ de X est

$$q_\alpha = \inf\{x \in \mathbb{R} : F(x) \geq \alpha\} \tag{3}$$

Si F est continue, $F(q_\alpha) = \alpha$ et si F est strictement croissante, q_α est unique.

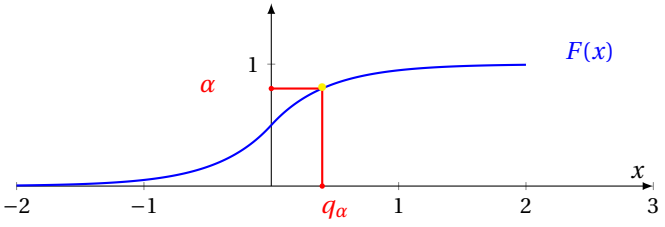

Figure 4.1 – Quantile et fonction de répartition

4.2 Méthodes pour déterminer une région de confiance

4.2.1 Fonction pivotale

Définition 68

Un pivot $Z = Z(X,\theta)$ est une fonction de la variable d'observation X et du paramètre θ telle que la loi de Z soit libre en θ.

On dit que θ est un paramètre de position pour T si la loi de $T-\theta$ ne dépend pas de θ, c'est à dire si $T-\theta$ est un pivot.

On dit que θ est un paramètre d'échelle pour T si la loi de $T \div \theta$ ne dépend pas de θ, c'est à dire si $T \div \theta$ est un pivot.

Le pivot se construit généralement à partir d'un estimateur de θ. Une fois déterminé, on cherche, grâce aux quantiles de la loi, un évènement B (qui ne dépend pas non plus de θ) tel que $\mathbb{P}_\theta[Z \in B] = 1-\alpha$. La région de confiance est alors de la forme $IC = [\theta \in \Theta : Z \in B]$.

Exemple:

Soit $X_1,...,X_n$ un n-échantillon de loi $\mathcal{N}(\theta,1)$ et soit $q = q_{1-\alpha/2}$ la quantile d'ordre $1-\alpha/2$ de la loi normale centrée réduite. Comme $\sqrt{n}(\overline{X}-\theta)$ est une v.a. pivotale de loi $\mathcal{N}(0,1)$, alors

$$\mathbb{P}_\theta\left[\sqrt{n}\left|\overline{X}-\theta\right| \leq q\right] = \Pi(q) - \Pi(-q) = 1-\alpha \tag{4}$$

où Π est la fonction de répartition de la loi normale centrée réduite, donnée par

$$\Pi(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-t^2/2}dt \tag{5}$$

et donc

$$\mathbb{P}_\theta\left(\theta \in \left[\overline{X} - \frac{q}{\sqrt{n}}; \overline{X} + \frac{q}{\sqrt{n}}\right] \leq q\right) = 1-\alpha \tag{6}$$

L'intervalle de confiance est

$$IC_{1-\alpha}(\theta) = \left[\overline{X} - \frac{q_{1-\alpha/2}}{\sqrt{n}} \; ; \; \overline{X} + \frac{q_{1-\alpha/2}}{\sqrt{n}}\right] \tag{7}$$

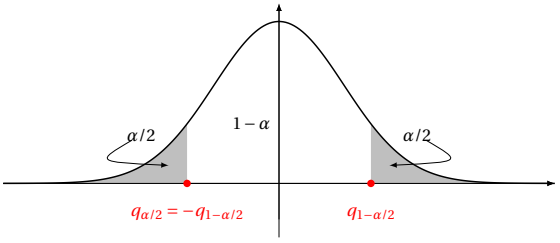

Figure 4.2 – Intervalle de confiance bilatéral de risque α.

Exemple:

Soit $(X_1,...,X_n)$ un n-échantillon de loi $\mathcal{N}(m,\sigma^2)$.

• Supposons $\theta = m$ inconnu et σ connu.

Un estimateur de θ est donnée par $\overline{X} \sim \mathcal{N}(\theta,\sigma^2/n)$. On peut alors choisir comme pivot

$$Z = \frac{\overline{X}-\theta}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1) \tag{8}$$

Soit $\alpha \in]0,1[$. Le choix des quantiles se fait en déterminant q_1, q_2 tels que $\mathbb{P}_\theta[q_1 \leq Z \leq q_2] = 1-\alpha$.

La symétrie de la loi $\mathcal{N}(0,1)$ permet de choisir un unique quantile q d'ordre $1-\alpha/2$ contenant le mode de la densité de la gaussienne pour avoir $\mathbb{P}_\theta[|Z| \leq q] = 1-\alpha$. L'intervalle de confiance est alors

$$IC_{1-\alpha}(\theta) = \left[\overline{X} - \frac{\sigma}{\sqrt{n}} q_{1-\alpha/2} \; ; \; \overline{X} + \frac{\sigma}{\sqrt{n}} q_{1-\alpha/2} \right] \tag{9}$$

• Supposons m et σ inconnu et posons $\theta = m$.

La v.a. $Z = (\overline{X} - \theta) \div (\sigma/\sqrt{n})$ définie précédemment suit une loi $\mathcal{N}(0,1)$. Posons

$$Y = \frac{n-1}{\sigma^2} S^2 \tag{10}$$

où S^2 est la variance empirique non biaisée. Y suit une loi χ^2_{n-1} et d'après le théorème de Cochran, $S^2 \perp\!\!\!\perp \overline{X}$. Nous savons en outre que

$$T = \frac{Z}{\sqrt{Y/(n-1)}} = \frac{\overline{X} - m}{\sqrt{S^2/n}} \tag{11}$$

suit une loi de Student à $n-1$ degrés de libertés, ne dépend que de m (mais sa loi n'en dépend pas, elle est libre de m) et peut donc être choisi comme pivot. Par symétrie de la loi de Student, le quantile t d'ordre $1-\alpha/2$ de cette loi de Student est tel que $-t$ est le quantile d'ordre $\alpha/2$. L'intervalle de confiance est donc

$$IC_{1-\alpha}(m) = \left[\overline{X} - \frac{S}{\sqrt{n}} t_{1-\alpha/2} \; ; \; \overline{X} + \frac{S}{\sqrt{n}} t_{1-\alpha/2} \right] \tag{12}$$

• Supposons m et σ inconnu et posons $\theta = (m, \sigma^2)$.

C'est une région de confiance inclut dans $\mathbb{R}^2$ que nous allons déterminer. Posons

$$Z = \left(\frac{\overline{X} - m}{\sigma/\sqrt{n}}, \frac{n-1}{\sigma^2} S^2 \right) \sim \mathcal{N}(0,1) \otimes \chi^2_{n-1} \tag{13}$$

Notons z le quantile d'ordre $1-\beta/2$ d'une loi normale centrée réduite et c_1, c_2 les quantiles d'ordre $\beta'/2$ et $1-\beta'/2$ d'une loi du chi deux à $n-1$ degrés de libertés. Les deux composantes de Z étant indépendantes,

$$\mathbb{P}_\theta\left(Z \in [-z,z] \times [c_1,c_2]\right) = \mathbb{P}_\theta\left(Z_1 \in [-z,z]\right) \times \mathbb{P}_\theta\left(Z_2 \in [c_1,c_2]\right) = (1-\beta)(1-\beta') = 1-\alpha \tag{14}$$

en posant $\alpha = (\beta + \beta') + \beta\beta'$. La région de confiance sera donc de la forme

$$RC_{1-\alpha}(\theta) = \left[\overline{X} - \frac{\sigma}{\sqrt{n}} z \; ; \; \overline{X} + \frac{\sigma}{\sqrt{n}} z \right] \times \left[\frac{n-1}{c_1} S^2 \; ; \; \frac{n-1}{c_2} S^2 \right] \tag{15}$$

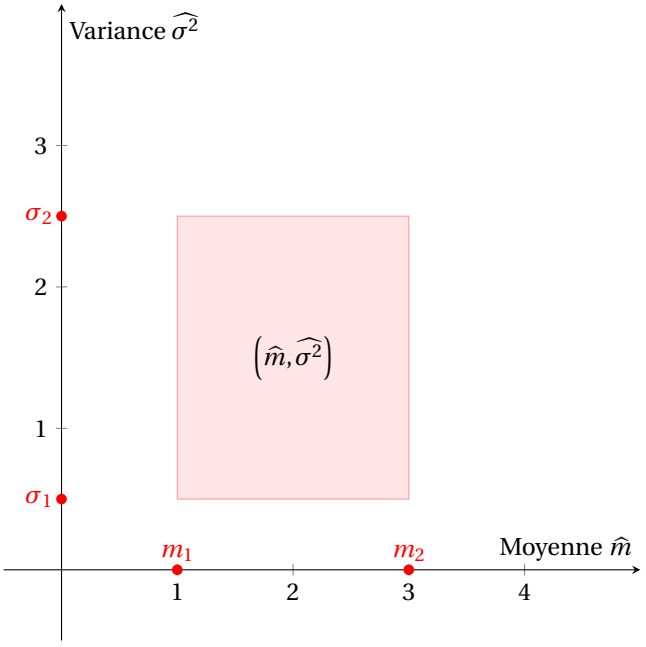

Figure 4.3 – Région de confiance de (m, σ^2) pour un échantillon gaussien.

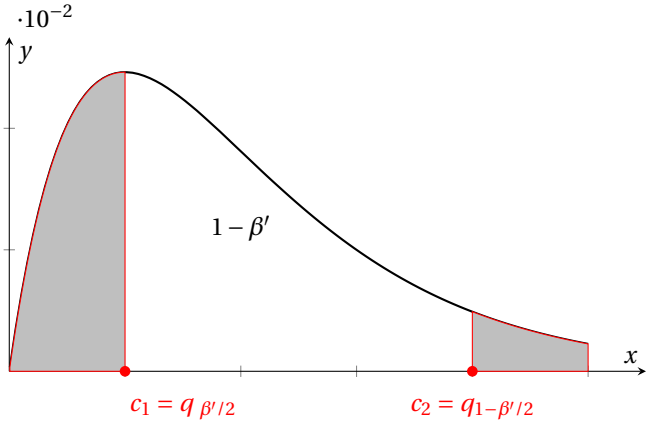

Figure 4.4 – Intervalle de confiance bilatéral (loi du khi deux) pour le paramètre σ^2 au risque β'.

4.2.2 Utilisation des inégalités de Tchebychev et de Hoeffding

Lorsque la v.a. est discrète ou que la loi n'est pas connue, on ne peut pas construire un intervalle de confiance comme vu précédemment, mais l'on peut introduire la notion plus faible d'intervalle de confiance par excès dans lequel

$$\mathbb{P}_\theta [\theta \in I] \geq 1-\alpha \tag{16}$$

Théorème 69 — Inégalité de Tchebychev

$$\mathbb{P}[|X - \mathbb{E}[X]| \geq \epsilon] \leq \frac{\mathbb{V}(X)}{\epsilon^2} \tag{17}$$

Démonstration

À faire à titre d'exercice (calculer $\mathbb{E}[|X|]$ sur l'évènement $[|X| \geq \epsilon]$ et sur son complémentaire).

□

Théorème 70 — Inégalité de Hoeffding

Soient $X_1, ..., X_n$ n v.a.i.i.d. telles que $a \leq X_1 \leq b$ $\mathbb{P}$-p.s. et $S_n = \sum_{i=1}^n X_i$. Alors pour tout $t > 0$,

$$\mathbb{P}\left[\left|\sum_{i=1}^n (X_i - \mathbb{E}[X_i])\right| \geq t\right] \leq 2\exp\left(-\frac{2t^2}{n(b-a)^2}\right) \tag{18}$$

soit encore

$$\mathbb{P}[|S_n - \mathbb{E}[S_n]| \geq t] \leq 2\exp\left(-\frac{2t^2}{n(b-a)^2}\right) \tag{19}$$

Démonstration

Cette inégalité de concentration très importante est proposée en exercice dans les travaux dirigés.

□

Exemple:

Considérons un n-échantillon $X_1, ..., X_n$ dont la loi, à support dans un intervalle $[a,b]$ indépendant de θ, admet un moment d'ordre 1. Cherchons un intervalle de confiance pour la moyenne commune $\theta = m$ des X_i.

L'inégalité de Tchebychev donne comme intervalle de confiance (par excès) de niveau $1-\alpha$

$$I_1 = \left[\overline{X} - \frac{b-a}{\sqrt{n\alpha}} \,;\, \overline{X} + \frac{b-a}{\sqrt{n\alpha}}\right] \tag{20}$$

L'inégalité de Hoeffding permet d'améliorer cet intervalle car

$$\mathbb{P}_\theta\left[|\overline{X}-\theta|\geq t\right]\leq 2\exp\left(-\frac{2nt^2}{(b-a)^2}\right) \tag{21}$$

En posant

$$t=(b-a)\sqrt{\frac{1}{2n}\ln\frac{2}{\alpha}} \tag{22}$$

on obtient un second intervalle (par excès également)

$$I_2=\left[\overline{X}-(b-a)\sqrt{\frac{1}{2n}\ln\frac{2}{\alpha}}\,;\,\overline{X}+(b-a)\sqrt{\frac{1}{2n}\ln\frac{2}{\alpha}}\right] \tag{23}$$

Les contributions de la taille de l'échantillon et de la longueur du support sont les mêmes, mais la contribution de α est nettement meilleure dans le second cas, lorsque α est proche de 0. En effet, quand α tend vers 0, le risque diminue et donc l'intervalle de confiance s'élargit. Mais il s'élargit beaucoup moins vite avec la seconde inégalité (à vitesse logarithmique) qu'avec la première (à la vitesse de $\sqrt{\alpha}$).

4.2.3 Inversion d'un test statistique

Les tests sont le sujet du paragraphe suivant. Nous y verrons le lien entre test et région de confiance et une méthode pour déterminer des intervalles de confiance à partir des régions de rejet des tests.

Pour la construction d'intervalles de confiance, il faut privilégier les pivots fonctions de statistiques exhaustives et complètes. En effet, le VUMSB est fonction d'une statistique exhaustive complète et la longueur de l'intervalle de confiance sera donc proportionnelle à la variance de l'estimateur considéré.

4.3 région de confiance asymptotique

4.3.1 Définition

Définition 71

Soit $X_1,...,X_n$ un n-échantillon de loi $\mathbb{P}_\theta$.

Un pivot asymptotique $Z_n=Z(X_1,...,X_n,\theta)$ est une suite de v.a. telle que Z_n converge en loi vers une v.a. qui ne dépend pas de θ.

4.3.2 Méthode du pivot asymptotique

Soit $\widehat{\theta}_n$ un estimateur $\sqrt{n}$-consistant et asymptotiquement normal de θ (ou plus généralement de $g(\theta)$) :

$$\sqrt{n}\left(\widehat{\theta}_n-\theta\right)\rightsquigarrow\mathcal{N}(0,\sigma^2) \tag{24}$$

Alors un pivot asymptotique est donné par

$$Z_n=\frac{\sqrt{n}\left(\widehat{\theta}_n-\theta\right)}{\sigma}\rightsquigarrow\mathcal{N}(0,1) \tag{25}$$

Si σ est connu, on calcule le quantile q d'ordre $1-\alpha/2$ de la loi normale centrée réduite et l'on a alors, par définition de la convergence en loi,

$$\begin{cases}\lim_{n\to+\infty}\mathbb{P}_\theta[Z_n\leq q]=1-\alpha/2\\ \lim_{n\to+\infty}\mathbb{P}_\theta[Z_n\leq -q]=\alpha/2\end{cases} \tag{26}$$

L'intervalle de confiance asymptotique de niveau $1-\alpha$ pour θ est alors donné par

$$IC_{1-\alpha}(\theta)=\left[\left|\frac{\widehat{\theta}_n-\theta}{\sigma/\sqrt{n}}\right|\leq q\right] \tag{27}$$

4.3.3 Méthode de Wald

Lorsque σ n'est pas connu ou dépend de θ, la méthode de Wald consiste à substituer dans la méthode précédente $\sigma(\theta)$ par un estimateur consistant $\widehat{\sigma}(\theta)$ de l'écart type et d'utiliser ensuite le lemme de Slutsky.

$$Z_n = \frac{\sqrt{n}\left(\widehat{\theta}_n - \theta\right)}{\widehat{\sigma}} \rightsquigarrow \frac{1}{\sigma(\theta)}\mathcal{N}(0, \sigma(\theta)^2) = \mathcal{N}(0,1) \tag{28}$$

Si q est le quantile d'ordre $1-\alpha/2$ d'une loi $\mathcal{N}(0,1)$, alors l'intervalle de confiance asymptotique est

$$IC_{1-\alpha}(\theta) = \left[\widehat{\theta} - q\frac{\widehat{\sigma}}{\sqrt{n}}\ ;\ \widehat{\theta} + q\frac{\widehat{\sigma}}{\sqrt{n}}\right] \tag{29}$$

Exemple:

On considère un n-échantillon de bernoulli de paramètre θ à estimer. Le paramètre d'intérêt est la moyenne de chaque v.a. et un estimateur est donné par la moyenne empirique. Le pivot asymptotique est donné par

$$\sqrt{\frac{n}{\overline{X}(1-\overline{X})}}(\overline{X} - \theta) \tag{30}$$

En effet, le théorème de la limite centrale montre que

$$\sqrt{n}\frac{\overline{X} - \theta}{\sqrt{\theta(1-\theta)}} \rightsquigarrow \mathcal{N}(0,1) \tag{31}$$

Et comme $\mathbb{E}[\overline{X}] = \theta$, en remplaçant θ par cet estimateur et en utilisant le lemme de Slutsky, le pivot asymptotique converge en loi vers une gaussienne centrée réduite. L'intervalle de confiance asymptotique est alors

$$IC_{1-\alpha}(\theta) = \left[\overline{X} - \frac{q}{\sqrt{n}}\sqrt{\overline{X}(1-\overline{X})}\ ;\ \overline{X} + \frac{q}{\sqrt{n}}\sqrt{\overline{X}(1-\overline{X})}\right] \tag{32}$$

où q est la quantile d'ordre $1-\alpha/2$ de la loi $\mathcal{N}(0,1)$.

4.3.4 Stabilisation de la variance

Soit $\Psi : \Theta \longrightarrow \mathbb{R}$ une fonction différentiable, telle que $\Psi'(\theta) = \nabla\Psi(\theta) = 1/\sigma(\theta)$. Si

$$\sqrt{n}\left(\widehat{\theta}_n - \theta\right) \rightsquigarrow \mathcal{N}(0, \sigma(\theta)^2) \tag{33}$$

alors par la méthode delta, on a

$$\sqrt{n}\left(\Psi(\widehat{\theta}_n) - \Psi(\theta)\right) \rightsquigarrow \mathcal{N}(0, J_\Psi(\theta)\sigma(\theta)^2 J'_\Psi(\theta)) = \mathcal{N}(0,1) \tag{34}$$

$\sqrt{n}\left(\Psi(\widehat{\theta}_n) - \Psi(\theta)\right)$ est donc un pivot asymptotique. Si Ψ est bijective et croissante et si q est le quantile d'ordre $1-\alpha/2$ d'une loi normale centrée réduite, alors on en déduit un intervalle de confiance asymptotique de la forme

$$IC_{1-\alpha}(\theta) = \left[\Psi^{-1}\left(\Psi(\widehat{\theta}) - \frac{q}{\sqrt{n}}\right); \Psi^{-1}\left(\Psi(\widehat{\theta}) + \frac{q}{\sqrt{n}}\right)\right] \tag{35}$$

4.3.5 Majoration de $\sigma(\theta)$

Soit $M > 0$ tel que $\sigma(\theta) < M$ pour tout θ. On a

$$1-\alpha = \lim_{n\to+\infty} \mathbb{P}_\theta\left[|\widehat{\theta} - \theta| \le q\frac{\sigma(\theta)}{\sqrt{n}}\right] \le \mathbb{P}_\theta\left[|\widehat{\theta} - \theta| \le q\frac{M}{\sqrt{n}}\right] = \mathbb{P}_\theta\left[\theta \in \left[\widehat{\theta} - q\frac{M}{\sqrt{n}}; \widehat{\theta} + q\frac{M}{\sqrt{n}}\right]\right] \tag{36}$$

ce qui donne un intervalle de confiance par excès de θ à un niveau $\ge 1-\alpha$.

4.3.6 Inversion d'un test statistique

Considérons un test asymptotique de région critique R relativement à un paramètre θ, pour un risque α fixé. Alors on peut considérer la région d'acceptation R comme l'intervalle de confiance au niveau $1-\alpha$ pour θ, grâce à l'équivalence

$$X \in \overline{R}_\theta \iff \theta \in \mathrm{IC}(X) \tag{37}$$

4.4 Choix des quantiles

Une loi de probabilité de fonction de répartition F est unimodale s'il existe x_M tel que F soit convexe avant x_M et concave après. x_M est alors le mode de F.

Propriété

Soit f la densité du pivot Z. Si f est unimodale de mode $_M$ et vérifie, pour $a \leq b$,

$$\bullet \mathbb{P}[a \leq Z \leq b] = \int_a^b f(x)dx = 1-\alpha \tag{38}$$

$$\bullet f(a) = f(b) > 0 \tag{39}$$

$$\bullet a < x_M < b \tag{40}$$

Alors $[a,b]$ est l'intervalle de longueur minimale vérifiant la première condition

Ainsi, par exemple, $q_{1-\alpha/2}$ est le meilleur choix de quantile pour une loi symétrique. $q_{\alpha/2}$ et $q_{1-\alpha/2}$ ne sont pas optimaux pour une loi du chi deux, mais sont souvent utilisés par défaut.

4.5 Intervalle de crédibilité dans un cadre bayésien

Dans un cadre bayésien où l'on suppose que $X \sim \mathbb{P}_\theta$ et $\theta \sim \Pi$, l'analogue d'un intervalle de confiance est appelé intervalle de crédibilité.

Définition 73

Un intervalle de crédibilité de niveau $1-\alpha$ pour θ est un sous-ensemble $IC_\Pi(X) \subset \Theta$ vérifiant

$$\mathbb{P}\left[H \in IC_\Pi(X) \mid X = x\right] = \mathbb{P}\left[\theta \in IC_\Pi(X) \mid X = x\right] = 1-\alpha \tag{41}$$

Lorsque θ est un paramètre réel, alors $IC_\Pi(X) = [a,b]$ avec a, b vérifiant

$$\mathbb{P}[a \leq \theta \leq b \mid X = x] = \int_a^b L(\theta, x)d\theta = 1-\alpha \tag{42}$$

Dans un intervalle de confiance, le paramètre est fixe et les bornes sont aléatoires (en tout cas tant que l'observation sur X n'a pas été réalisée), tandis que dans un intervalle de crédibilité, les bornes sont fixes tandis que le paramètre est aléatoire.

Exemple:

Considérons un n-échantillon gaussien de loi $\mathcal{N}(\theta, \sigma^2)$ où σ^2 est connu et θ a pour loi *a priori* également gaussienne $\mathcal{N}(a, b^2)$. Alors on a vu que la loi *a posteriori* était une gaussienne $\mathcal{N}(m, \tau^2)$ avec

$$\begin{cases} \dfrac{1}{\tau^2} = \dfrac{1}{b^2} + \dfrac{n}{\sigma^2} \\ \dfrac{m}{\tau^2} = \dfrac{a}{b^2} + \dfrac{n\overline{x}}{\sigma^2} \end{cases} \tag{43}$$

soit encore

$$\begin{cases} m = \dfrac{\sigma^2/n}{\sigma^2/n + b^2} a + \dfrac{b^2}{\sigma^2/n + b^2} \overline{x} \\ \tau^2 = \dfrac{\sigma^2 b^2/n}{\sigma^2/n + b^2} \end{cases} \tag{44}$$

Sachant $[X = x]$, la v.a. $Z = (\theta - m)/\tau \sim \mathcal{N}(0,1)$. Pour un risque $\alpha \in [0,1]$, si q est le quantile d'ordre $1-\alpha/2$ de la loi $\mathcal{N}(0,1)$, alors l'intervalle

$$IC_\Pi(X) = \left[m - q\tau \,;\, m + q\tau\right] \tag{45}$$

est $(1-\alpha)$ crédible car $\mathbb{P}_{Z|[X=x]}[-q \leq Z \leq q] = 1-\alpha$.

Chapitre 5

Tests statistiques

5.1 Tests paramétriques

5.1.1 Hypothèse nulle et hypothèse alternative

On dispose d'un n-échantillon $X = (X_1, ..., X_n)$ et l'on veut construire une fonction de décision ϕ (appelée également fonction de test) à valeurs binaire, de la forme

$$\phi(X) = \mathbb{1}_{[T(X) \in R]} \tag{1}$$

$R \subset \mathbb{X}^n$ est la région de rejet du test et $T(X)$ est une statistique des observations X. On identifie souvent la fonction de décision et la zone de rejet.

Définition 74

Un test pur est une statistique $\phi : \mathbb{X} \longrightarrow \{0,1\}$

$R = \phi^{-1}(1) = \{x \in \mathbb{X} : \ \phi(x) = 1\}$ est la région de rejet du test.

$\overline{R} = \phi^{-1}(0) = \{x \in \mathbb{X} : \ \phi(x) = 0\}$ est la région d'acceptation du test.

On suppose que l'on se place dans le cas d'un modèle paramétrique où le paramètre $\theta \in \Theta$. Θ est partitionné en deux sous-ensembles Θ_0 et Θ_1 de telle sorte que $\Theta_0 \cap \Theta_1 = \emptyset$. $\theta_0 \in \Theta_0$, $\theta_1 \in \Theta_1$.

Définition 75

Faire une hyopthèse $\mathcal{H}_0$ sur θ, appelée hypothèse nulle, consiste à se donner un sous-ensemble Θ_0 de Θ. On note cette hypothèse:

$$\mathcal{H}_0 : \ \theta \in \Theta_0 \tag{2}$$

On note $\mathcal{H}_1$ l'hypothèse alternative:

$$\mathcal{H}_1 : \ \theta \in \Theta_1 \tag{3}$$

où $\Theta_1 \subset \Theta$ et $\Theta_0 \cap \Theta_1 = \emptyset$ (souvent $\overline{\Theta_0} = \Theta_1$).

Définition 76

L'erreur de première espèce consiste à rejeter $\mathcal{H}_0$ à tort.

L'erreur de seconde espère consiste à ne pas rejeter $\mathcal{H}_0$ à tort.

Dans l'approche de Neyman-Pearson, les deux hypothèses ne sont pas symétriques: on part de l'idée que $\mathcal{H}_0$ est vraie *a priori* et l'on va essayer de se convaincre (ou pas) à l'aide des observations, que ce n'est pas le cas. $\mathcal{H}_0$ est l'hyopthèse privilégiée, en ce sens qu'elle est supposée vérifiée tant que les observations ne conduisent

pas à la rejeter. Les observations amènent alors à accepter $\mathscr{H}_0$ (on dit plutôt « ne pas rejeter ») ou bien à rejeter cette hypothèse au profit de $\mathscr{H}_1$.

Il existe différentes formes d'hypothèses:

• Hypothèses simples: $\Theta_0 = \{\theta_0\}$ et $\Theta_1 = \{\theta_1\}$ sont des singletons.

• Hypothèse simple contre hypothèse composite:
Test bilatère: $\Theta_0 = \{\theta_0\}$ et $\Theta_1 \neq \{\theta_0\}$ ou test unilatère: $\Theta_0 = \{\theta_0\}$ et $\Theta_1 = \{\theta \in \Theta : \theta > \theta_0\}$.

• Hypothèses composites: lorsque les deux ensembles Θ_0 et Θ_1 contiennent au moins deux éléments.

5.1.2 Risque, niveau, puissance, efficacité

Le risque de première espèce (fausse alerte) s'écrit

$$\alpha(\theta) = \mathbb{P}_\theta(R) = \mathbb{P}_\theta\left[\phi(X) = 1\right] = \mathbb{E}_\theta[\phi(X)] \text{ , pour } \theta \in \Theta_O \tag{4}$$

Le risque de seconde espèce (non détection d'une erreur) s'écrit

$$\beta(\theta) = \mathbb{P}_\theta\left(\overline{R}\right) = \mathbb{P}_\theta\left[\phi(X) = 0\right] = \mathbb{E}_\theta[1 - \phi(X)] \text{ , pour } \theta \in \Theta_1 \tag{5}$$

α est donc une fonction défini sur Θ_0 et à valeurs dans $[0,1]$, tandis que β est une fonction définie sur Θ_1 à valeurs dans $[0,1]$.

On souhaiterait minimiser ces deux risques simultanément mais ce n'est pas possible et il faut faire un compromis (minimiser α revient à minimiser la taille de R, tandis que minimiser β revient à augmenter la taille de R). On peut privilégier le risque de fausse alerte et demander qu'il soit aussi petit que possible.

Définition 77

Le niveau α^* d'un test est

$$\alpha^* = \sup_{\theta \in \Theta_0} \alpha(\theta) \tag{6}$$

Ce supremum existe (il est défini sur un ensemble non vide et borné) mais n'est pas toujours atteint. On dira que le test est de niveau α si $\alpha = \alpha^*$ et qu'il est de seuil α si $\alpha < \alpha^*$.

$$\mathbb{P}_{\theta_0}(R) \leq \alpha^* \tag{7}$$

Pour tout test, on dispose donc d'une famille de régions critiques $(R_\alpha)_{\alpha \in]0,1[}$ indicées par le niveau α du test.

Supposons qu'il existe une statistique à valeurs réelles $T(X)$ (appelée statistique de test) telle que

$$\phi = \mathbb{1}_{[s,\infty[}(T(X)) = \mathbb{1}_{[T(x) \geq s]} \tag{8}$$

Définition 78

La valeur critique (ou p-valeur) d'un tel test ϕ est, pour toute observation x donnée,

$$p(x) = \sup_{\theta \in \Theta_0} \mathbb{P}_\theta[T(X) > T(x)] \tag{9}$$

$p(x)$ est le plus petit niveau à partir duquel on rejète $\mathscr{H}_0$ sur la base de l'observation x.

$$\phi(x) = 1 \iff T(x) > s \iff p(x) < \alpha \tag{10}$$

Dans le cas où $\Theta = \{\theta_0\}$, la p-valeur est donc la v.a. fonction de X égale à la plus petite valeur de α à partir de laquelle on rejète $\mathscr{H}_0$ au vue des observations. C'est à dire:

• $\forall \alpha < p(x)$, on ne rejète pas $\mathscr{H}_0$.
• $\forall \alpha > p(x)$, on rejète $\mathscr{H}_0$.

Interprétation de la p-valeur: on calcule la probabilité (sous $\mathscr{H}_0$) d'obtenir pour T une observation aussi extrême que $T(x)$. Cette probabilité est la p-valeur. Avec la p-valeur, on cherche à quantifier le fait d'être bon ou pas dans le fait de rejeter $\mathscr{H}_0$. La p-valeur représente donc le point de basculement entre rejet et non-rejet, pour une observation donnée. Si elle est petite, on peut rejeter $\mathscr{H}_0$. Nous allons revenir rapidement sur cette définition dans les exemples ci-après.

On peut démontrer (à titre d'exercice) que $p(X)$ suit une loi uniforme sur $[0,1]$.

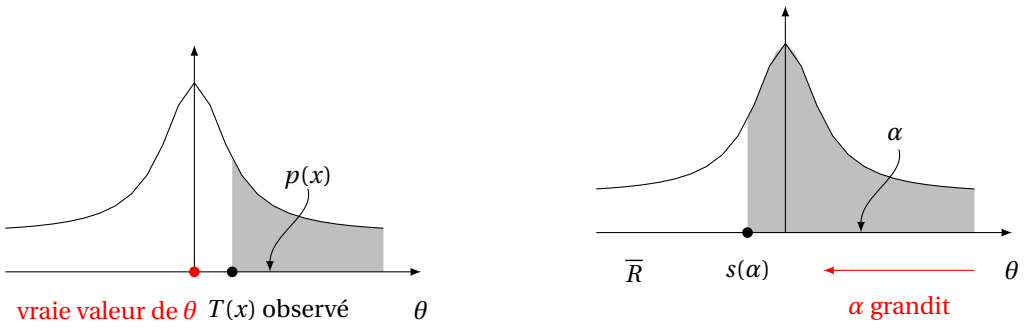

Figure 5.1 – p -valeur et risque

À partir du risque de seconde espèce β, on calcule la puissance γ du test:

$$\gamma(\theta) = 1 - \mathbb{P}_\theta[T = 0] = 1 - \beta(\theta) \tag{11}$$

de sorte que γ prolonge à Θ_1 la fonction α.

$$\alpha(\theta_0) = \mathbb{P}_{\theta_0}[T = 1] = \mathbb{E}_{\theta_0}[T] \tag{12}$$

$$\gamma(\theta_1) = \mathbb{P}_{\theta_1}[T = 1] = \mathbb{E}_{\theta_1}[T] \tag{13}$$

Et l'on peut résumer les liens entre α, β et γ dans le tableau ci-dessous:

Décision \ Réalité	$\mathscr{H}_0$	$\mathscr{H}_1$
$\mathscr{H}_0$	$1-\alpha$	β
$\mathscr{H}_1$	α	$\gamma = 1-\beta$

Définition 79

La courbe d'efficacité d'un test T est la fonction

$$h : \Theta_0 \cup \Theta_1 \longrightarrow [0,1] : \theta \mapsto h(\theta) = 1 - \gamma(\theta) \tag{14}$$

Nous donnons maintenant plusieurs exemples pour illustrer toutes ces définitions.

Exemple:

Le ministère de la santé étudie régulièrement la nécessité de prendre des mesures contre la consommation d'alcool et l'efficacité de ces mesures. L'Insee fournit à cet effet des données annuelles de consommation moyenne d'alcool par personne et par jour. En 1991, la loi Évin interdit la publicité sur les boissons alcoolisées et lance une campagne de sensibilisation sous forme de spots publicitaires. Avant la loi Évin, la consommation d'alcool moyenne chez les personnes de plus de 15 ans était de 35 g par jour. L'objectif premier de la loi était de baisser cette consommation journalière moyenne à 33 g. En se basant sur l'observation des consommations moyennes de 1991 à 1994, le ministère a fixé la règle de décision suivante: si la moyenne des consommations journalières sur ces quatre années est supérieures à 34.2 g, alors les mesures prises ont été inefficaces.

On suppose que les données recueillies sont des réalisations de v.a. supposées i.i.d. et de loi normale $\mathcal{N}(\theta, \sigma^2)$ avec $\sigma = 2$. On note X_i, $i = 1...4$, les moyennes de consommation journalières pour chacune des quatre années de l'enquête et l'on note $\overline{X}$ la moyenne empirique des X_i.

Le test est donc le suivant:

$$\begin{cases} \mathcal{H}_0 : \theta = 33 \\ \mathcal{H}_1 : \theta = 35 \end{cases} \tag{15}$$

L'hypothèse nulle est $\theta = 33$ car on souhaite que la politique ait été efficace. On suppose donc l'effet significatif *a priori*.

Il semble naturel de choisir $\overline{X}$ comme statistique pour mesurer l'effet moyen de la consommation journalière (pourquoi ?). Nous commençons par calculer l'erreur de première espèce α du test, en notant $s = 34.2$ le seuil de significativité choisi (arbitrairement) :

$$\alpha = \mathbb{P}_{33}[\overline{X} > s] = \mathbb{P}_{33}[\overline{X} - 33 > s - 33] = 1 - \Pi(34.2 - 33) \approx 0,1151 \tag{16}$$

Où Π est la fonction de répartition d'une loi normale centrée réduite:

$$\Pi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt \tag{17}$$

$\overline{X}$ suit une loi normale de moyenne θ inconnue et de variance 1, car la somme de v.a. normales indépendantes est également normale et nous avons choisi l'écart type afin que la variance de $\overline{X}$ vaille 1.

Le risque de seconde espèce est

$$\beta = \mathbb{P}_{35}[\overline{X} \leq s] = \mathbb{P}_{35}[\overline{X} - 35 \leq s - 35] = \Pi(s - 35) \approx 0,212 \tag{18}$$

Les deux hypothèses jouent des rôles dissymétriques car les conséquences liées aux deux types d'erreurs ne sont pas les mêmes. Lorsque la valeur du seuil s augmente, il est clair que α décroit vers 0 tandis que β croit vers 1. On ne donc pas minimiser les deux simultanément (on remarque par ailleurs que $\alpha + \beta \neq 1$).

La valeur de α par défaut est souvent 5%. Le seuil correspondant, que l'on notera $s^* = s(\alpha)$ (ou parfois $c^* = c(\alpha)$) vérifie :

$$\mathbb{P}_{33}\left[\overline{X} > s^*\right] = 0,05 = \alpha \iff 1 - \mathbb{P}_{33}\left[\overline{X} - 33 < s^* - 33\right] = 0,05 \tag{19}$$

$$\iff \mathbb{P}\left[Z < s^* - 33\right] = 0,95 \tag{20}$$

$$\iff \Pi(s^* - 33) = \Pi(1,65) \tag{21}$$

$$\iff s^* = 33 + 1,65 = 34,65 \tag{22}$$

Ainsi, si l'on pose $s = 34.65$, le test défini par $\phi(x) = \mathbb{1}_{[\overline{x} > 34.65]}$ est donc de niveau $\alpha = 5\%$.

Si l'on a observé $x = (34.7\ ,\ 34.4\ ,\ 33.7\ ,\ 33.3)$, alors la moyenne est $\overline{x} = 34.025$ et la p-valeur du test (pour cette observation) est

$$p(x) = \mathbb{P}_{33}[\overline{X} \leq 34.025] \approx 0.153 \tag{23}$$

0.153 est le niveau minimum auquel on va rejeter $\mathcal{H}_0$ pour ces observations. 0.05 est plus petit que cette p-valeur et effectivement, on accepte $\mathcal{H}_0$ à ce niveau de 5% car la moyenne observée de 34.025 est bien inférieure à 34.645.

Exemple:

On considère un n-échantillon gaussien de loi $\mathcal{N}(\theta, 1)$ dont la moyenne θ est inconnue. On définit le test :

$$\begin{cases} \mathcal{H}_0 : \theta \leq \theta_0 \\ \mathcal{H}_1 : \theta > \theta_0 \end{cases} \tag{24}$$

Où $\theta \in \mathbb{R}$ est fixé. Une idée naturelle est de considéter la moyenne empirique $\overline{X} \sim \mathcal{N}(\theta, 1/n)$. En ce cas, et sous l'hypothèse $\mathcal{H}_0$, la v.a. $\sqrt{n}(\overline{X} - \theta)$ suit une loi $\mathcal{N}(0,1)$.

Une région de rejet serait de la forme $R(q) = \left[x : \sqrt{n}(\overline{X} - \theta_0) > q\right]$. Le quantile q doit être déterminé pour que le test soit de niveau α. On considère donc, sous l'hypothèse $\mathcal{H}_0$, un $\theta \leq \theta_0$:

$$\mathbb{P}_\theta[R(q)] = \mathbb{P}_\theta\left[\sqrt{n}\left(\overline{X} - \theta_0\right) > q\right] = \mathbb{P}_\theta\left[\sqrt{n}\left(\overline{X} - \theta\right) + \sqrt{n}(\theta - \theta_0) > q\right] \tag{25}$$

$$\leq \mathbb{P}_\theta\left[\sqrt{n}(\overline{X} - \theta) > q\right] \tag{26}$$

$$\leq \mathbb{P}[Z > q] \tag{27}$$

L'égalité n'est obtenue que si $\theta = \theta_0$. Ainsi, $\alpha^* = \sup_{\theta\in\Theta_0} \mathbb{P}_\theta[R(q)] = \mathbb{P}[Z > q] = 1 - \Pi(q)$

où Π est la fonction de répartition de la loi normale centrée réduite et $Z \sim \mathcal{N}(0,1)$. q doit donc être égal au quantile $q_{1-\alpha}$ d'une loi normale centrée réduite pour que le test soit de niveau α.

Exemple:

On effectue n lancers à pile ou face indépendants avec une même pièce.

On souhaite savoir si la pièce est équilibrée. On dispose donc d'un n-échantillon de Bernoulli de paramètre θ et le test est défini par

$$\begin{cases} \mathcal{H}_0 : \theta = 1/2 \\ \mathcal{H}_1 : \theta \neq 1/2 \end{cases} \tag{28}$$

Nous avons $\theta = \mathbb{E}_\theta[X_1]$ et $n\overline{X}$ suit une loi binomiale $B(n,\theta)$. Sous l'hypothèse $\mathcal{H}_0$, cette v.a. suit donc une loi $B(n,1/2)$. On propose comme région de rejet

$$R(a,b) = \left[b \leq \overline{x} \text{ ou } \overline{x} \leq a\right] \tag{29}$$

avec $a < 1/2 < b$. Il faut également calculer a et b de telle sorte que $\mathbb{P}_{[\theta=1/2]}[R(a,b)] = \alpha$. On a

$$\mathbb{P}_{1/2}\left(\left[\overline{X} < a\right] \cup \left[\overline{X} > b\right]\right) = \mathbb{P}_{1/2}\left(\left[n\overline{X} < na\right] \cup \left[n\overline{X} > nb\right]\right) = \mathbb{P}_{1/2}\left[\overline{X} < a\right] + \mathbb{P}_{1/2}\left[\overline{X} > b\right] = \alpha \tag{30}$$

chacune des deux probabilités ne peut prendre qu'un nombre fini de valeurs car $n\overline{X}$ est à valeurs dans $\{0,...,n\}$. On ne peut pas choisir α quelconque: il n'existe pas toujours de test de niveau α. Par contre, on peut toujours définir un seuil. Dans l'exemple, on peut montrer (à titre d'exercice) que si $n = 1000$, $a = 0,469$ et $b = 0,531$ donne un test de seuil $\alpha = 5\%$. C'est d'ailleurs le test le plus puissant parmi tous les tests dont le milieu de $[a,b]$ est $1/2$.

Test pur et test aléatoire

Il existe une notion de test aléatoire (ou test stochastique ou encore test mixte) qui étend la notion de test « pur ». Dans un test pur, il n'y a que deux décisions possibles: on rejète ou on ne rejète par l'hypothèse. L'exemple discret précédent montre que ce type de test peut poser problème. Pour éviter cela, on peut considérer que la valeur $T(x)$ peut être comprise entre 0 et 1 et représenter une probabilité de rejeter $\mathcal{H}_0$ avec l'observation x. L'ensemble $T^{-1}(]0,1[)$ s'appelle alors région d'hésitation du test.

L'intérêt des tests aléatoires est essentiellement théorique: ils permettent de démontrer des théorèmes d'existence et d'unicité. Une interprétation d'un test aléatoire consiste à définir, à x fixé, $\phi(x)$ comme étant la probabilité de rejeter $\mathcal{H}_0$. On considère alors un tirage aléatoire Y suivant une loi de Bernoulli de paramètre $\phi(x)$. Soit y une réalisation de ce tirage. On rejète $\mathcal{H}_0$ si $y = 1$ et on accepte si $y = 0$.

La loi de Y sachant $[X = x]$ est une loi de Bernoulli de paramètre $\phi(x)$. Ainsi,

$$\mathbb{E}[Y|X = x] = \phi(x) \tag{31}$$
$$\mathbb{E}[Y|X] = \phi(X) \tag{32}$$
$$\mathbb{E}[Y] = \mathbb{E}[\phi(X)] \tag{33}$$
$$\alpha(\theta) = \mathbb{E}_\theta[\phi(X)] \tag{34}$$
$$\beta(\theta) = \mathbb{E}_\theta[1 - \phi(X)] \tag{35}$$
$$\tag{36}$$

Construction pratique des tests

L'idée de Neyman et Pearson (en 1933) est d'avoir une approche en deux temps: on commence par contrôler le risque de première espèce α, puis une fois ce risque fixé, on cherche parmi les tests restants ceux qui minimisent β. Pour construire un test, on procède suivant les étapes suivantes:

1. Détermination du modèle statistique représentant le phénomène aléatoire étudié.
2. Détermination des hypothèses $\mathcal{H}_0$ et $\mathcal{H}_1$.
3. Détermination d'une statistique de test T.
4. Détermination de la forme de la fonction de test ϕ et de la forme de la région critique.

5. Détermination des constantes et quantiles intervenant dans ϕ de telle façon que le test soit de niveau ou de seuil donné. Cela revient à déterminer la frontière de la région critique.

6. Conclusion au vu de l'observations x.

7. Évaluation de la puissance du test.

5.1.3 Tests de deux hypothèses simples

On parle de test à hypothèses simples lorsque $\mathscr{H}_i$ est un singleton, c'est à dire que le test est de la forme $\theta = \theta_0$ vs $\theta = \theta_1$

Le lemme de Neyman-Pearson

Un test est dit « uniformément plus puissant de niveau α » et noté UPPα s'il est de niveau α et qu'il est plus puissant que tout autre test de niveau α. Il est donc optimal parmi tous les tests lorsque le niveau α est fixé. Dans le cas d'hypothèses simples, il existe toujours un test UPPα et ce test (indiqué par une $*$) vérifie

$$\gamma^*(\theta_1) = \mathbb{P}_{\theta_1}(R^*) = \sup_R \mathbb{P}_{\theta_1}(R) = \sup_R \gamma(\theta_1) \tag{37}$$

Le sup étant pris sur l'ensemble des tests de niveau α (donc également sur l'ensemble des régions de rejet possibles, puisqu'un test pur est caractérisé par sa région de rejet). Pour montrer l'existence d'un test UPPα, on pose

$$\phi(x) = \phi^*(x) + \frac{\alpha - \alpha^*}{1 - \alpha^*}(1 - \phi^*(x)) \tag{38}$$

Donc $\phi(x) \in [0,1]$ et ϕ définit bien un test. Par ailleurs, $\mathbb{E}_{\theta_0}[\phi] = \alpha$ et comme $\phi^*(x) \leq \phi(x)$, pour tout $\theta_1 \in \Theta_1$ on a $\mathbb{E}_{\theta_1}[\phi^*] \leq \mathbb{E}_{\theta_1}[\phi]$. Par conséquent, ϕ est un test de niveau α uniformément plus puissant que ϕ^*.

Définition 80

Un test est sans biais si sa puissance γ est supérieure ou égale à α:

$$\gamma(\theta_1) \geq \alpha \tag{39}$$

Propriété

Soit $\alpha \in]0,1[$. Un test UPPα est sans biais.

Démonstration

Soit ϕ^* un test UPPα. On considère le test suivant (appelé test constant) : $\phi(x) = \alpha$ rejète $\mathscr{H}_0$ quelque soit les observations. Ce test est de niveau α et donc $\mathbb{E}_{\theta_1}[\phi^*] \geq \mathbb{E}_{\theta_1}[\phi] = \alpha$.

□

Supposons l'existence d'une mesure dominante μ telle que $\forall \theta \in \Theta$, on puisse définir une vraisemblance $L(X,\theta)$.

Définition 82

Le rapport de vraisemblance $LR(X,\theta_1,\theta_0)$ est la v.a. définie par

$$LR(X,\theta_1,\theta_0) = \frac{L(X,\theta_1)}{L(X,\theta_0)} \tag{40}$$

Elle est définie si $L(X,\theta_0) > 0$ $\mathbb{P}_\theta$ presque sûrement. En ce cas, un test de Neyman-Pearson est un test dont la région critique est de la forme

$$R(c) = [LR(X) > c] \tag{41}$$

Dans le cas contraire, la région de rejet peut s'écrire sous la forme $R(c) = [L(X,\theta_1) > cL(X,\theta_0)]$

Lemme de Neyman-Pearson

Soit $\alpha \in]0,1[$ fixé.

- Un test de Neyman-Pearson de niveau $\alpha \in]0,1[$ est UPPα.
- Tous les tests UPPα sont de cette forme pour une constante $c(\alpha)$ donnée.

$$\mathbb{P}_{\theta_0}[LR(X,\theta_1,\theta_0) > c(\alpha)] = \alpha \tag{42}$$

- La fonction de test est de la forme

$$\phi(x) = \begin{cases} 1 & \text{si} \quad LR(x,\theta_1,\theta_0) > c \\ k & \text{si} \quad LR(x,\theta_1,\theta_0) = c \\ 0 & \text{si} \quad LR(x,\theta_1,\theta_0) < c \end{cases} \tag{43}$$

Si $\mathbb{P}_{\theta_0}[LR(x,\theta_1,\theta_0) = c] = 0$, alors $k = 0$ ou 1 et le test est pur. C'est le cas lorsque les v.a. sous jacentes sont à densité continue.

L'idée de Neyman-Person est donc de fixer α, puis ensuite de minimiser β. Ceci est équivalent à fixer α et à maximiser la puissance γ pour cette valeur de α donnée. Un test UPPα est un test pour lequel α a été fixé et l'on cherche le test de plus grande puissance.

Démonstration

idée: si la v.a. $LR(X)$ a une fonction de répartition continue, alors c existe. Il existe une version plus générale avec des tests stochastiques qui assure que $c(\alpha)$ existe toujours. La démonstration est un peu longue et nous l'omettons donc.

□

On dit qu'un test est consistant lorsque $\lim_{\theta\to+\infty}\gamma(\theta) = 1$.

Exemple:

Considérons un n-échantillon de loi normale $X \sim \mathcal{N}(\theta,\sigma^2)$. σ^2 est connu et le test est défini par $\mathcal{H}_0 : \theta_0 = 0$ vs $\mathcal{H}_1 : \theta_1 = 1$

Un calcul rapide du quotient des deux vraisemblance montre que la région de rejet est donnée par

$$R = [LR(X) > c(\alpha)] = \left[\exp\left(\frac{n}{2\sigma^2}\left(\overline{X}-1\right)\right) > c(\alpha)\right] = \left[\overline{X} > s(\alpha)\right] \tag{44}$$

où $s(\alpha)$ est une fonction de n et de σ. On pourrait déterminer $s(\alpha)$ en fonction de $c(\alpha)$ mais cela n'a aucune utilité; il suffit seulement de savoir que la région de rejet est de la forme $\overline{X}$ supérieur à un seuil. On choisit alors ce seuil $s = s(\alpha)$ de telle sorte que

$$\mathbb{P}_{\theta_0}[R] = \mathbb{P}_{\theta_0}\left[\overline{X} > s\right] = \mathbb{P}\left[Z > \frac{\sqrt{n}}{\sigma}s\right] = \alpha \tag{45}$$

où $Z \sim \mathcal{N}(0,1)$. En effet, sous $\mathcal{H}_0$, $\overline{X} \sim \mathcal{N}(0,\sigma^2/n)$. On veut donc

$$1 - \Pi\left(\frac{\sqrt{n}s}{\sigma}\right) = \alpha \iff s = s_\alpha = \frac{\sigma}{\sqrt{n}}\phi^{-1}(1-\alpha) = \frac{\sigma}{\sqrt{n}}q_{1-\alpha} \tag{46}$$

où $q = q_{1-\alpha}$ est le quantile d'ordre $1-\alpha$ de la loi normale centrée réduite.

Finalement, le test de Neyman-Pearson est défini par

$$\phi(X) = \mathbb{1}_{[X\in R]} = \mathbb{1}_{\left[\overline{X} > q\sigma/\sqrt{n}\right]} \tag{47}$$

Quelle est la puissance du test ?

$$\gamma = \mathbb{P}_{\theta_1}[R] = \mathbb{P}_1\left(\frac{\sqrt{n}\,\overline{X}}{\sigma} > q_{1-\alpha}\right) = \mathbb{P}_{\theta_1}\left[\frac{\sqrt{n}(\overline{X}-1)}{\sigma} + \frac{\sqrt{n}}{\sigma} > q_{1-\alpha}\right] \tag{48}$$

$$= 1 - \Pi\left(q - \frac{\sqrt{n}}{\sigma}\right) \overset{n\to+\infty}{\longrightarrow} 1 \tag{49}$$

Le test est donc consistant.

Les observations ayant conduits à la valeur $\overline{x}$ utile dans le test, si α petit, $\overline{x} < d_\alpha$ et si α grand, $\overline{x} > d_\alpha$, de sorte que le moment de basculement intervient quand α vérifie $\overline{x} = d_\alpha$. La p-valeur vaut ici

$$p(x) = \alpha(\overline{x}) = \mathbb{P}_0\left[\overline{X} > d_{\alpha(\overline{x})}\right] = \alpha \tag{50}$$

La faiblesse dans la notion de p-valeur est le fait qu'elle ne prenne en compte que l'hypothèse nulle.

5.1.4 Modèles à rapport de vraisemblance monotone

Définition 84

Un modèle statistique $\left(\mathbb{X}, \mathfrak{B}, (\mathbb{P}_\theta)_{\theta\in\Theta}\right)$ avec $\Theta \subset \mathbb{R}$ est dit à rapport de vraisemblance monotone en une statistique T si, pour $\theta_0 < \theta_1$, il existe une fonction $\psi : \mathbb{R} \longrightarrow \overline{\mathbb{R}}$ monotone, dépendant de θ_0 et θ_1, telle que

$$LR(x, \theta_1, \theta_0) = \psi(T(x)) \tag{51}$$

On peut également parler de modèle à rapport de vraisemblance strictement monotone, croissant ou décroissant en T, si ψ l'est.

L'exemple fondamental de modèle à rapport de vraisemblance monotone est est donné par la famille des modèles exponentiels, pour lesquels le rapport de vraisemblance se factorise au travers de la statistique privilégiée.

$$LR(x, \theta', \theta) = \exp\left[T(x)(\lambda(\theta') - \lambda(\theta)) - (\beta(\theta') - \beta(\theta))\right] \tag{52}$$

Le test se lit alors directement sur la statistique T.

5.1.5 Tests composites

Extension du lemme de Neyman-Pearson

Dans un test composite, Θ_0 ou Θ_1 contiennent plus d'un élément.

$$\begin{cases} \mathcal{H}_0 : \theta \in \Theta_0 \\ \mathcal{H}_1 : \theta \in \Theta_1 \\ \Theta_0 \cap \Theta_1 = \emptyset \end{cases} \tag{53}$$

Les définitions sont les mêmes que pour les tests simples, mais dans le cas composite, les tests UPPα n'existent que très rarement. Malgré tout, certains résultats du lemme de Neyman-Pearson s'étendent au cas composite.

Lemme de Neyman-Pearson étendu

Soit $\alpha \in]0,1[$ fixé.

S'il existe $\theta_0 \in \Theta_0$ tel que $\mathbb{E}_{\theta_0}[T] = \alpha$ et si pour tout $\theta_1 \in \Theta_1$ on peut trouver une constante $c(\alpha)$ vérifiant

$$T(x) = \mathbb{1}_{[LR(x,\theta_1,\theta_0) > s(\alpha)]} \tag{54}$$

Alors T est UPPα parmi les tests de niveau α.

Exemple:

Considérons un n-échantillon gaussien $X_i \sim \mathcal{N}(\theta, \sigma^2)$ et le test défini par

$$\begin{cases} \mathcal{H}_0 : \theta \leq \theta_0 \\ \mathcal{H}_1 : \theta > \theta_0 \end{cases} \tag{55}$$

Calculons la puissance et le niveau du test. Soit $\theta \in \Theta_0$:

$$\gamma(\theta) = \mathbb{P}_\theta\left[\overline{X} > \frac{\sigma}{\sqrt{n}} q_{1-\alpha}\right] = \mathbb{P}_\theta\left[\frac{\sqrt{n}}{\sigma}(\overline{X}-\theta) + \frac{\sqrt{n}}{\sigma}(\theta-\theta_0) > \frac{\sigma}{\sqrt{n}} q_{1-\alpha}\right] = 1 - \Pi\left(q_{1-\alpha} - \frac{\sqrt{n}}{\sigma}(\theta-\theta_0)\right) \tag{56}$$

$$= \Pi\left(\frac{\sqrt{n}}{\sigma}(\theta-\theta_0) - q_{1-\alpha}\right) \tag{57}$$

La fonction de répartition Π est continue et croissante, par conséquent:

$$\alpha^* = \sup_{\theta \leq \theta_0} \mathbb{P}_\theta[R] = \Pi\left(-q_{1-\alpha}\right) = 1 - (1-\alpha) \tag{58}$$

Le test est donc bien de niveau α. Considérons maintenant un autre test de niveau α et de région critique R' (pour les mêmes hypothèses). La région critique R_α dans le test de Neyman-Pearson de $\mathscr{H}_0 : \theta = \theta_0$ vs $\mathscr{H}_1' : \theta = \theta_1$ vérifie, d'après le lemme de Neyman-Pearson,

$$\mathbb{P}_{\theta_1}(R_\alpha) \geq \mathbb{P}_{\theta_1}(R_\alpha) \tag{59}$$

et le test est donc bien UPPα. Pour un test composite, ceci se produira rarement.

Tests unilatères de $\mathscr{H}_0 : \theta = \theta_0$ vs $\mathscr{H}_1 : \theta > \theta_0$ ou $\mathscr{H}_1 : \theta < \theta_0$

Dans le cas d'un modèle statistique à rapport de vraisemblance strictement croissant en $T(x)$, il existe pour tout $\alpha \in]0,1[$, un test UPPα de la forme

$$\phi(x) = \mathbb{1}_{[T(x)>s]} \tag{60}$$

Ceci est une conséquence du lemme de Neyman-Pearson étendu. La même conclusion est valable pour une hypothèse $\mathscr{H}_1 : \theta < \theta_0$ avec

$$\phi(x) = \mathbb{1}_{[T(x)<s]} \tag{61}$$

Tests unilatères de $\mathscr{H}_0 : \theta \leq \theta_0$ vs $\mathscr{H}_1 : \theta > \theta_0$

Théorème 86 de Lehmann

Soit un modèle statistique à rapport de vraisemblance strictement croissant en $T(x)$. Pour tout $\alpha \in]0,1[$, il existe un test de niveau α UPPα de la forme

$$\phi(x) = \mathbb{1}_{[T(x)>s]} \tag{62}$$

De plus, le niveau α est atteint pour $\theta = \theta_0$

Tests bilatères $\mathscr{H}_0 : \theta \leq \theta_1$ ou $\theta \geq \theta_2$ vs $\mathscr{H}_1 : \theta \in]\theta_1, \theta_2[$

Il existe un résultat dans le cas où le modèle est exponentiel, dominé par une mesure μ, de statistique privilégiée ϕ, de paramètre $\lambda(\theta)$ et de fonction de log-partition β. Si λ est strictement croissante, de telle sorte que le modèle soit à rapport de vraisemblance strictement croissant en $T(x)$, alors pour tout $\alpha \in]0,1[$, il existe un test de niveau α UPPα de la forme

$$\phi(x) = \mathbb{1}_{[c_1<\phi(x)<c_2]} \tag{63}$$

De plus, le niveau est atteint en θ_1 ou θ_2.

Il existe d'autres résultats du même type avec les tests bilatères, mais il n'est pas question ici d'en donner de liste exhaustive. Nous en verrons certains exemples en travaux dirigés.

5.1.6 Tests asymptotiques

Test de Wald unidimensionnel

On dispose d'un estimateur T_n de θ déterminé à partir du n-échantillon.

Définition 87

Un test de région de rejet R est de niveau asymptotique α si

$$\sup_{\theta\in\Theta_0}\lim_{n\to+\infty}\mathbb{P}_\theta(R)=\alpha \tag{64}$$

Théorème 88 Test de Wald en dimension 1

On suppose que $\Theta\subset\mathbb{R}$ et que les deux hypothèses suivantes sont vérifiées:

$$\sqrt{n}(T_n-\theta)\rightsquigarrow\mathcal{N}(0,\sigma^2(\theta)) \tag{65}$$

$$\sigma^2:\Theta\longrightarrow]0,\infty[\text{ est continue} \tag{66}$$

Alors le test asymptotique $\mathcal{H}_0:\theta=\theta_0$ vs $\mathcal{H}_1:\theta\neq\theta_0$ de région de rejet

$$R=\left[\left|\sqrt{n}\frac{T_n-\theta_0}{\sigma(T_n)}\right|>q\right] \tag{67}$$

où q est le quantile d'ordre $1-\alpha/2$ d'une loi $\mathcal{N}(0,1)$, est un test de niveau α qui est consistant.

Démonstration:

Si $Z_n\rightsquigarrow Z$ avec $Z\mathcal{N}(0,1)$, alors

$$\lim_{n\to+\infty}\mathbb{P}(Z_n\in[a,b])=\mathbb{P}(Z\in[a,b]) \tag{68}$$

car la frontière de l'intervalle $[a,b]$ est de mesure nulle. Ainsi,

$$\lim_{n\to+\infty}\mathbb{P}_{\theta_0}(R)=\lim_{n\to+\infty}\mathbb{P}_{\theta_0}\left[\left|\sqrt{n}\frac{T_n-\theta_0}{\sigma(T_n)}\right|>q\right] \tag{69}$$

$$=\mathbb{P}[Z>q]=\alpha \tag{70}$$

Pour démontrer la constistance du test, nous devons montrer que $\forall\theta\in\Theta_1$, $\lim_n\mathbb{P}_\theta(R)=1$. Ceci revient à démontrer que la limite de la probabilité du complémentaire tend vers 0. Nous avons

$$\overline{R}=\left[\left|\sqrt{n}\frac{T_n-\theta_0}{\sigma(T_n)}\right|\leq q\right] \tag{71}$$

En utilisant l'inégalité $|x-y|\geq|x|-|y|$, il vient:

$$\mathbb{P}_\theta(\overline{R})=\mathbb{P}_\theta\left[\left|\sqrt{n}\frac{T_n-\theta}{\sigma(T_n)}+\sqrt{n}\frac{\theta-\theta_0}{\sigma(T_n)}\right|\leq q\right] \tag{72}$$

$$\leq\mathbb{P}_\theta\left[\sqrt{n}\frac{|\theta-\theta_0|}{\sigma(T_n)}-\sqrt{n}\frac{|T_n-\theta|}{\sigma(T_n)}\leq q\right] \tag{73}$$

$$\leq\mathbb{P}_\theta\left[|\theta-\theta_0|\leq\frac{\sigma(T_n)}{\sqrt{n}}\times\frac{\sqrt{n}}{\sigma(T_n)}|T_n-\theta|+\frac{\sigma(T_n)}{\sqrt{n}}q\right] \tag{74}$$

Le terme $\sigma(T_n)/\sqrt{n}$ tend vers 0 en probabilité quand n tend vers l'infini, tandis que $\sqrt{n}|T_n-\theta|/\sigma(T_n)$ converge en loi vers une gaussienne centrée réduite. L'expression à droite de l'inégalité tend donc vers 0 en loi et en probabilité et par suite, la probabilité tend vers 0 avec n.

□

Exemple:

On considère un modèle d'échantillonnage avec comme statistique $T_n=\widehat{\theta}_n^{\text{MV}}=\widehat{\theta}_n$ l'estimateur du maximum de vraisemblance de θ. On suppose que $\Theta\subset\mathbb{R}$ et que le modèle est dominé de vraisemblance $f(x,\theta)$.

Sous les conditions de régularités suffisantes, nous savons que

$$\sqrt{n}\left(\widehat{\theta}_n-\theta\right)\rightsquigarrow\mathcal{N}\left(0,\mathbb{I}(\theta)^{-1}\right) \tag{75}$$

L'application $\theta\longrightarrow\mathbb{I}(\theta)$ est continue et définie positive (ici, c'est donc un scalaire >0). Le test de Wald avec $\mathcal{H}_0:\theta=\theta_0$ *vs* $\mathcal{H}_0:\theta\neq\theta_0$ a pour région de rejet

$$R=\left[\left|\sqrt{n}\frac{\widehat{\theta}_n-\theta_0}{\sigma(\widehat{\theta}_n)}\right|>q\right] \tag{76}$$

Test de Wald multidimensionnel

On considère le test

$$\begin{cases} \mathcal{H}_0 : \theta \in \Theta_0 = \{\theta \in \mathbb{R}^d : g(\theta) = 0\} \\ \mathcal{H}_1 : \theta \notin \Theta_0 \end{cases} \tag{77}$$

où $g : \mathbb{R}^d \longrightarrow \mathbb{R}^n$ où $n \leq d$.

Théorème 89 — Test de Wald multidimensionnel

On suppose que les hypothèses suivantes sont vérifiées:

- $\sqrt{n}\left(\widehat{\theta}_n - \theta\right) \rightsquigarrow \mathcal{N}\left(0, \mathbb{I}(\theta)^{-1}\right)$
- $\mathbb{I}(\theta)^{-1}$ est une fonction continue de θ
- g est de classe C^1 sur $\mathbb{R}^d$ et sa matrice jacobienne $J_g(\theta)$ est continue en θ
- $\text{rg}\left(J_g(\theta)\right) = r$

Alors le test $\mathcal{H}_0 : g(\theta) = 0$ *vs* $\mathcal{H}_1 : g(\theta) \neq 0$ est de niveau asymptotique α et sa région de rejet est

$$R = \left[n g(\widehat{\theta})' B(\widehat{\theta})^{-1} g(\widehat{\theta}) > q \right] \tag{78}$$

Test du rapport de vraisemblance maximal

Considérons le test suivant

$$\begin{cases} \mathcal{H}_0 : \theta = \theta_0 \\ \mathcal{H}_1 : \theta \neq \theta_0 \end{cases} \tag{79}$$

Le théorème de Neyman Pearson ne s'applique pas car le test est composite. On suppose que $L(X,\theta)$ est continue en θ et l'on pose

$$\Lambda_n = \frac{\sup_{\theta \neq \theta_0} L(X,\theta)}{\sup_{\theta = \theta_0} L(X,\theta)} = \frac{\sup_{\theta \in \Theta} L(X,\theta)}{\sup_{\theta = \theta_0} L(X,\theta)} = \frac{L(X,\widehat{\theta}_n)}{L(X,\theta_0)} \tag{80}$$

Définition 90

Un test de vraisemblance maximale est de la forme

$$\phi(X) = \mathbb{1}_{[\Lambda_n(x) > s]} \tag{81}$$

Exemple:

Considèrons un n-échantillon gaussien $(X_1, ..., X_n)$ de $X \sim \mathcal{N}(\theta, \sigma^2)$ où σ^2 connu. Considérons le test composite $\mathcal{H}_0 : \theta = \theta_0$ vs $\mathcal{H}_1 : \theta \neq \theta_0$. L'estimateur du maximum de vraisemblance est $\widehat{\theta} = \overline{X}$. L'estimateur restreint est, par construction de l'hypothèse nulle, $\widehat{\theta}_0 = \theta_0$. La vraisemblance de l'échantillon est

$$L(x,\theta) = \prod_{i=1}^{n} \left[(2\pi\sigma^2)^{-1/2} \exp\left(-\frac{1}{2\sigma^2}(x_i - \theta)^2 \right) \right] \tag{82}$$

On pose alors

$$R = \left[LR(X, \widehat{\theta}, \widehat{\theta}_0) > s \right] = \left[\exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left((X_i - \overline{X})^2 - (X_i - \theta_0)^2 \right) \right) > s \right] \tag{83}$$

$$= \left[\left| \overline{X} - \theta_0 \right| > \delta \right] \tag{84}$$

En développant, en simplifiant, en composant par la réciproque d'exponentielle, etc.

Exemple:

Considérons un n-échantillon de loi de Poisson de paramètre λ et le test

$$\begin{cases} \mathscr{H}_0 : \lambda = \lambda_0 \\ \mathscr{H}_1 : \lambda \neq \lambda_0 \end{cases} \tag{85}$$

Il n'existe pas de test UPPα correspondant à ce problème. La vraisemblance de l'échantillon est

$$L(x,\lambda) = \lambda^{\sum_{i=1}^n x_i} \frac{e^{-n\lambda}}{\prod_{i=1}^n x_i!} \tag{86}$$

Nous savons que le maximum de vraisemblance est atteint pour $\widehat{\lambda} = \overline{x}$ et le maximum restreint est ici égal à λ_0 d'après la forme de l'hypothèse nulle. Le rapport de vraisemblance maximale est

$$\Lambda(x) = \left(\frac{\lambda_0}{\overline{x}}\right)^{\sum_{i=1}^n x_i} \times e^{n(\overline{x}-\lambda_0)} \tag{87}$$

On a donc

$$-2\ln\Lambda(x) = 2n\left(\overline{x}\ln(\overline{x}/\lambda_0) + \lambda_0 - \overline{x}\right) \tag{88}$$

Ici, $d = 1$ et $p = 0$. Le théorème de Wilks assure que, sous $\mathscr{H}_0$, lorsque n tend vers l'infini,

$$2n\left(\overline{X}\ln\left(\frac{\overline{X}}{\lambda_0}\right) + \lambda_0 - \overline{X}\right) \rightsquigarrow \chi^2_1 \tag{89}$$

La région de rejet sera donc de la forme

$$R = \left[2n\left(\overline{X}\ln\left(\frac{\overline{X}}{\lambda_0}\right) + \lambda_0 - \overline{X}\right) \geq t\right] \tag{90}$$

où t est le quantile adéquat de la loi du χ^2.

On fait à nouveau l'hypothèse que $\sqrt{n}(\widehat{\theta}_n - \theta)$ converge en loi vers $\mathscr{N}\left(0, \mathbb{I}(\theta)^{-1}\right)$ et que la matrice d'information $\mathbb{I}(\theta)$ est définie, inversible et intégrable. On note $H = (h_{ij})_{i,j}$ la matrice hessienne de la log-vraisemblance et l'on suppose en outre que

$$\forall x,\ \forall \theta, \theta' \in \Theta,\ |h_{ij}(x,\theta) - h_{ij}(x,\theta')| \leq M(x) \times ||\theta - \theta'||_2 \tag{91}$$

où M est une fonction mesurable vérifiant $\mathbb{E}[M(X_1)] < \infty$. On rappelle enfin que

$$h_{ij}(x,\theta) = \frac{\partial^2}{\partial\theta_i\theta_j} \ln f(x,\theta) \tag{92}$$

Théorème 91 — Théorème de Wilks (1938)

Sous les conditions précédentes, si $\Theta \subset \mathbb{R}^d$ est un ouvert convexe, alors

$$\lambda_n = 2\ln T_n \rightsquigarrow \chi^2_d \tag{93}$$

Idée de la démonstration

□

Si $\Theta = \mathbb{R}^d$ et si Θ_0 est un sous-espace vectoriel de dimension p, alors $2\ln\Lambda_n(X)$ converge en loi, sous $\mathscr{H}_0$, vers une loi χ^2_{d-p} (sous certaines hypothèses de régularité du modèle).

5.2 Quelques tests non paramétriques

5.2.1 Quelques exemples

5.2.2 Le test de Student

5.2.3 Le test de signe d'Arbuthnol

5.2.4 Statistiques de rangs

Chapitre 6

L, Z, U, M-estimateurs

Ce chapitre n'est pas au programme de l'examen.

Chapitre 7

Notion de robustesse

Ce chapitre n'est pas au programme de l'examen.

Chapitre 8

Applications

Annexe A

Quelques rappels utiles

A.1 Les notations $\mathbb{P}_\theta$, $\mathbb{E}_\theta$ et $\mathbb{V}_\theta$

L'espace probabilisé $(\Omega, \mathfrak{F}, \mathbb{P})$ dans lequel est défini l'expérience aléatoire initiale ne sert souvent à rien... La plupart du temps, les variables aléatoires utilisées sont des fonctions mesurables de Ω dans $\mathbb{X}$ et on peut supposer $\Omega = \mathbb{X}$ en effectuant une « identification canonique ». On notera $\mathbb{P}$ (avec une double barre) la probabilité (canonique) sous-jacente.

Dans un modèle paramétrique identifiable, chaque mesure de probabilité $\mathbb{P}_\theta$ de la famille est caractérisée par une valeur θ du paramètre. Lorsque l'on notre $X \sim \mathbb{P}_\theta$ on suppose que la loi de la v.a. X est égale à cette mesure. θ étant généralement inconnu, $\mathbb{P}_\theta$ l'est aussi.

Les notations $\mathbb{E}_\theta[X]$ et $\mathbb{V}_\theta(X)$ signifient que l'on calcule l'espérance et la variance de X en supposant que cette v.a. soit de loi $\mathbb{P}_\theta$, autrement dit,

$$\mathbb{E}_\theta[X] = \int_{\mathbb{X}} x \, \mathbb{P}_\theta(dx) \tag{1}$$

de façon plus générale, si h est une fonction continue bornée, l'expression $\mathbb{E}_\theta[h(X)]$ signifie

$$\mathbb{E}_\theta[h(X)] = \int_{\mathbb{X}} h(x) \mathbb{P}_\theta(dx) \tag{2}$$

On dit que $\mathbb{E}_\theta$ est l'espérance « sous θ » et $\mathbb{V}_\theta$ la variance « sous θ ».

A.2 Rappels sur les convergences de v.a.

Soit $(X_n)_{n\geq 0}$ une suite de variables aléatoires sur un espace probabilisé $(\Omega, \mathfrak{F}, \mathbb{P})$, à valeurs dans $\mathbb{R}^d$.

Définition 92

On dit que X_n converge vers X quand n tend vers l'infini

- $\mathbb{P}$-presque sûrement, et l'on note $X_n \overset{\mathbb{P}-\text{p.s.}}{\longrightarrow} X$ si, et seulement si, $\mathbb{P}\left[\lim_{n\to+\infty} X_n = X\right] = 1$
- en probabilité, et l'on note $X_n \overset{\mathbb{P}}{\longrightarrow} X$ si, et seulement si, $\forall \epsilon > 0$, $\lim_{n\to+\infty} \mathbb{P}\left[||X_n - X|| > \epsilon\right] = 0$
- en norme L^p, et l'on note $X_n \overset{L^p}{\longrightarrow} X$ si, et seulement si, $\lim_{n\to+\infty} ||X_n - X||_p = 0$
- en loi, et l'on note $X_n \rightsquigarrow X$ si, et seulement si, pour toute fonction f continue et bornée, $\lim_{n\to+\infty} \mathbb{E}[f(X_n)] = \mathbb{E}[f(X)]$

Propriété

$$X_n \overset{\mathbb{P}-\text{p.s.}}{\longrightarrow} X \iff \forall \epsilon > 0, \ \mathbb{P}\left(\overline{\lim_{n\to+\infty}} \left[||X_n - X|| > \epsilon\right]\right) = 0 \tag{3}$$

$$\iff \exists\, N \subset \Omega : \mathbb{P}(N) = 0 \text{ et } \forall \omega \notin N, \ \lim_{n\to+\infty} X_n(\omega) = X(\omega) \tag{4}$$

On rappelle que la limite supérieure d'une suite d'évènements $(A_n)_n$ est définie par

$$\overline{\lim_{n\to+\infty}} A_n = \limsup A_n = \bigcap_{n\geq 0} \bigcup_{k\geq n} A_k \tag{5}$$

Un élément $\omega \in \Omega$ appartient à l'évènement $\limsup A_n$ si, et seulement si, ω se trouve dans une infinité de A_n (c'est à dire qu'il existe une sous-suite n_k vérifiant $\omega \in A_{n_k}$ pour tout k).

Propriété

Soit F_n la fonction de répartition de X_n et soit F la fonction de répartition de X. Alors

$$X_n \rightsquigarrow X \iff \forall t \text{ où } F_n \text{ est } C^0,\ \lim_{n\to+\infty} F_n(t) = F(t) \tag{6}$$

Théorème 95 de Paul Lévy

Soit ϕ_n la fonction caractéristique de X_n et soit ϕ la fonction caractéristique de X. Alors

$$X_n \rightsquigarrow X \iff \forall t \in \mathbb{R},\ \lim_{n\to+\infty} \phi_n(t) = \phi(t) \tag{7}$$

Propriété

$$X_n \xrightarrow{\mathbb{P}-\text{p.s.}} X \Rightarrow X_n \xrightarrow{\mathbb{P}} X \tag{8}$$

$$X_n \xrightarrow{L^p} X \Rightarrow X_n \xrightarrow{\mathbb{P}} X \tag{9}$$

$$X_n \xrightarrow{\mathbb{P}} X \Rightarrow X_n \rightsquigarrow X \tag{10}$$

$$\forall q > p,\ X_n \xrightarrow{L^q} X \Rightarrow X_n \xrightarrow{L^p} X \tag{11}$$

Théorème 97 de Portmanteau

Pour toute suite de v.a. $(X_n)_n$ à valeurs réelles et pour toute v.a. X réelle, on a:

$$X_n \rightsquigarrow X \tag{12}$$

$$\iff \lim_{n\to+\infty} \mathbb{P}[X_n \leq x] = \mathbb{P}[X \leq x],\ \text{en tout point de continuité } x \text{ de la fonction } x \mapsto \mathbb{P}[X \leq x] \tag{13}$$

$$\iff \lim_{n\to+\infty} \mathbb{E}[f(X_n)] = \mathbb{E}[f(X)],\ \forall f \text{ continue et bornée} \tag{14}$$

$$\iff \lim_{n\to+\infty} \mathbb{E}[f(X_n)] = \mathbb{E}[f(X)],\ \forall f \text{ bornée et lipschitzienne} \tag{15}$$

$$\iff \liminf \mathbb{E}[f(X_n)] \geq \mathbb{E}[f(X)],\ \forall f \text{ continue et positive} \tag{16}$$

$$\iff \liminf \mathbb{P}[X_n \in O] \geq \mathbb{P}[X \in O],\ \forall\ O \text{ ouvert} \tag{17}$$

$$\iff \limsup \mathbb{P}[X_n \in F] \leq \mathbb{P}[X \in F],\ \forall\ F \text{ ouvert} \tag{18}$$

$$\iff \lim_{n\to+\infty} \mathbb{P}[X_n \in B] = \mathbb{P}[X \in B],\ \forall\ B \text{ borélien dont la frontière est de mesure nulle} \tag{19}$$

Théorème 98 de l'application continue

Soit $\psi : \mathbb{R}^k \longrightarrow \mathbb{R}^p$ une application continue presque sûrement. Alors

$$X_n \xrightarrow{\mathbb{P}-\text{p.s.}} X \Rightarrow \psi(X_n) \xrightarrow{\mathbb{P}-\text{p.s.}} \psi(X) \tag{20}$$

$$X_n \xrightarrow{\mathbb{P}} X \Rightarrow \psi(X_n) \xrightarrow{\mathbb{P}} \psi(X) \tag{21}$$

$$X_n \rightsquigarrow X \Rightarrow \psi(X_n) \rightsquigarrow \psi(X) \tag{22}$$

Théorème 99 de Prohorov

Pour toute suite de v.a. $(X_n)_n$ dans $\mathbb{R}^p$,

- Si $X_n \rightsquigarrow X$ alors la suite $(X_n)_n$ est uniformément tendue.
- Si $(X_n)_n$ est uniformément tendue, alors il existe une sous-suite $(n_k)_k$ telle que $X_{n_k} \rightsquigarrow X$

Propriété

Si c est une v.a. presque sûrement constante,

$$X_n \xrightarrow{\mathbb{P}} c \iff X_n \rightsquigarrow c \tag{23}$$

$$X_n \rightsquigarrow X \text{ et } Y_n \xrightarrow{\mathbb{P}} c \Rightarrow (X_n, Y_n) \rightsquigarrow (X, c) \tag{24}$$

$$X_n \xrightarrow{\mathbb{P}} X \text{ et } Y_n \xrightarrow{\mathbb{P}} Y \Rightarrow (X_n, Y_n) \xrightarrow{\mathbb{P}} (X, Y) \tag{25}$$

$$\text{Si } X_n \perp\!\!\!\perp Y_n \forall n \text{ et } X \perp\!\!\!\perp Y, \text{ alors : } X_n \rightsquigarrow X \text{ et } Y_n \rightsquigarrow Y \Rightarrow (X_n, Y_n) \rightsquigarrow (X, Y) \tag{26}$$

La convergence presque sûre, la convergence en norme L^p et la convergence en probabilité sont stables par addition, multiplication et division. Ceci est faux pour la convergence en loi, mais dans le cas où l'une des suites converge vers une constante, on dispose du lemme (très utile) suivant:

Lemme de Slutsky

Si $X_n \rightsquigarrow X$ et $Y_n \rightsquigarrow c$ où c est une constante, alors:

- $X_n + Y_n \rightsquigarrow X + c$
- $Y_n X_n \rightsquigarrow cX$
- $X_n / Y_n \rightsquigarrow X/c$ si $c \neq 0$

A.3 Les o et $\mathcal{O}$ stochastiques

Définition 102

$$X_n = o_{\mathbb{P}}(R_n) \iff X_n = Y_n R_n \text{ avec } Y_n \xrightarrow{\mathbb{P}} 0 \tag{27}$$

$$X_n = \mathcal{O}_{\mathbb{P}}(R_n) \iff X_n = Y_n R_n \text{ avec } Y_n \text{ bornée en probabilité} \tag{28}$$

Propriété

$$o_{\mathbb{P}}(1) + o_{\mathbb{P}}(1) = o_{\mathbb{P}}(1) \tag{29}$$

$$o_{\mathbb{P}}(1) + \mathcal{O}_{\mathbb{P}}(1) = \mathcal{O}_{\mathbb{P}}(1) \tag{30}$$

$$o_{\mathbb{P}}(1) \times \mathcal{O}_{\mathbb{P}}(1) = o_{\mathbb{P}}(1) \tag{31}$$

$$o_{\mathbb{P}}(X_n) = X_n o_{\mathbb{P}}(1) \tag{32}$$

$$\mathcal{O}_{\mathbb{P}}(X_n) = X_n \mathcal{O}_{\mathbb{P}}(1) \tag{33}$$

$$o_{\mathbb{P}}(\mathcal{O}_{\mathbb{P}}(1)) = o_{\mathbb{P}}(1) \tag{34}$$

Propriété

Soit R une fonction définie sur $\mathbb{R}^k$, nulle en 0. Soit X_n une suite de v.a. à valeurs dans le domaine de définition de R. Pour tout $p>0$, lorsque x tend vers 0,

- Si $R(x)=o(||x||^p)$, alors $R(X_n)=o_{\mathbb{P}}(||x||^p)$
- Si $R(x)=\mathcal{O}(||x||^p)$, alors $R(X_n)=\mathcal{O}_{\mathbb{P}}(||x||^p)$

A.4 Lois usuelles

Nom	paramètres	$\mathbb{P}[X=k]$	$\mathbb{E}[X]$	$\mathbb{V}(X)$	À valeurs dans
Bernoulli	$p\in[0,1]$	$\mathbb{P}[X=1]=p$ $\mathbb{P}[X=0]=1-p=q$	p	pq	$\{0,1\}$
Équiprobable		$\mathbb{P}[X=k]=1/n$	$\frac{n+1}{2}$	$\frac{n^2-1}{12}$	$[\![1...n]\!]$
Binomiale	$n\in\mathbb{N}^*$ et $p\in[0,1]$	$\mathbb{P}[X=k]=C_n^k p^k q^{n-k}$	np	npq	$[\![0...n]\!]$
Géométrique	$p\in[0,1]$	$\mathbb{P}[X=k]=q^{k-1}p$	$\frac{1}{p}$	$\frac{q}{p^2}$	$\mathbb{N}^*$
Poisson	$\lambda>0$	$\mathbb{P}[X=k]=e^{-\lambda}\frac{\lambda^k}{k!}$	λ	λ	$\mathbb{N}$
Binomiale négative	$n\in\mathbb{N}^*$ et $p\in[0,1]$	$\mathbb{P}[X=k]=C_{k-1}^{n-1}p^n q^{k-n}$	$\frac{n}{p}$	$\frac{nq}{p^2}$	$k\geq n$

- Une v.a. de loi binomiale $\mathscr{B}(n,p)$ est la somme de n v.a. de Bernoulli indépendantes de paramètre p.
- Une v.a. de loi binomiale négative $\mathscr{B}(n,p)$ est la somme de n v.a. géométriques indépendantes de paramètre p. Attention à cette loi qui connaît plusieurs définitions différentes.
- Les lois binomiales, de Poisson et binomiales négatives sont stables par additions indépendantes.

Nom	paramètres	densité $f(x)$	$\mathbb{E}[X]$	$\mathbb{V}(X)$	À valeurs dans
Loi uniforme	α,β	$\frac{1}{\beta-\alpha}\times\mathbb{1}_{[\alpha,\beta]}(x)$	$\frac{\alpha+\beta}{2}$	$\frac{(\beta-\alpha)^2}{12}$	$[\alpha,\beta]$
Loi exponentielle VF	λ	$\lambda e^{-\lambda x}\mathbb{1}_{[0,+\infty[}(x)$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	$[0,+\infty[$
Loi exponentielle VO	θ	$\frac{1}{\theta}e^{-x/\theta}\mathbb{1}_{[0,+\infty[}(x)$	θ	θ^2	$[0,+\infty[$
Loi de Laplace		$\frac{1}{2}e^{-\lvert x\rvert}$	0	2	$\mathbb{R}$
Loi de Cauchy		$\frac{1}{\pi}\frac{1}{1+x^2}$	∞	∞	$\mathbb{R}$
Loi triangulaire		$(1-\lvert x\rvert)\mathbb{1}_{[-1,1]}(x)$	0	1/6	$[-1,1]$
Loi normale	$m\in\mathbb{R},\sigma>0$	$\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2\right)$	m	σ^2	$\mathbb{R}$
Loi gamma	$a>0,\lambda>0$	$\frac{x^{a-1}\lambda^a}{\Gamma(a)}e^{-\lambda x}\mathbb{1}_{[0,\infty[}(x)$	$\frac{a}{\lambda}$	$\frac{a}{\lambda^2}$	$[0,\infty[$
Loi béta	$a>0,b>0$	$\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}x^{a-1}(1-x)^{b-1}\mathbb{1}_{[0,1]}(x)$	$\frac{a}{a+b}$	$\frac{ab}{(a+b)^2(a+b+1)}$	$[0,1]$

A.5 Index des notations