UNIVERSITÉ DE NANTES

Master 1 Ingénierie Statistique

Statistique Inférentielle

Paul Rochet

Contents

1 Introduction **2**
1.1 Echantillonnage . . . 2
1.2 Paramètres et estimateurs . . . 3
1.3 Statistiques empiriques . . . 3
1.4 Statistiques d'ordre . . . 4

2 Modèles paramétriques **5**
2.1 Vraisemblance . . . 5
2.2 Familles exponentielles . . . 7
2.3 Vecteurs Gaussiens . . . 8

3 Information et exhaustivité **10**
3.1 Information de Fisher . . . 10
3.2 Conditionnement et Exhaustivité . . . 11
3.3 Efficacité . . . 15

4 Estimation ponctuelle **19**
4.1 Méthode des moments . . . 19
4.2 Maximum de vraisemblance . . . 22

5 Estimation par région de confiance **26**
5.1 Définitions générales . . . 26
5.2 Le modèle Gaussien . . . 27

6 Tests paramétriques **29**
6.1 Définitions générales . . . 29
6.2 Test du rapport de vraisemblance . . . 32

1 Introduction

1.1 Echantillonnage

On dispose de données $x = (x_1, ..., x_n)$ réelles ($x_i \in \mathbb{R}, \forall i$) ou vectorielles ($x_i \in \mathbb{R}^k, \forall i$) issues d'une expérience (mesures d'une quantité physique, sondage, etc...). Dans le cadre de l'inférence statistique, on va supposer que les données x sont la réalisation d'une variable aléatoire de loi inconnue que l'on cherche à connaître. Pour rendre la tâche possible (que peut-on apprendre sur une variable aléatoire à partir d'une seule réalisation?), le *modèle d'échantillonnage* fait l'hypothèse supplémentaire que les x_i sont des réalisations de variables aléatoires indépendantes et de même loi μ. On se placera toujours dans ce cadre sauf mention contraire. On dira que le vecteur $X = (X_1, ..., X_n)$ est un échantillon de variables indépendantes et identiquement distribuées (iid).

Remarque 1.1. *On distingue l'échantillon $(X_1, ..., X_n)$ qui est aléatoire et les données $(x_1, ..., x_n) = (X_1(\omega), ..., X_n(\omega))$ qui sont une réalisation de l'échantillon. Lors d'une étude statistique, on observe seulement l'image de ω par les fonctions $X_1, ..., X_n$.*

Contrairement aux statistiques descriptives qui consistent essentiellement à résumer l'information contenue dans des données à l'aide d'indicateurs (moyenne, médiane, variance etc...) ou de représentations graphiques (diagrammes, boxplots, histogrammes etc...), la statistique inférentielle cherche à fournir un modèle probabiliste qui explique les données observées. Les observations sont donc traitées comme des réalisations de variables aléatoires de loi inconnue, dont on cherche à retrouver certaines caractéristiques (ex: espérance, probabilité d'un événement, densité etc...).

La statistique inférentielle permet de répondre à différentes questions:

1. Le choix d'un modèle permettant de décrire le phénomène aléatoire
2. L'estimation de la loi des observations ou de paramètres de cette loi
3. La construction d'intervalles de confiances
4. La vérification d'hypothèses par des tests

Plus le modèle est précis, plus il sera facile d'évaluer la loi des observations. Par exemple, si on dispose d'un échantillon $X_1, ..., X_n$ issu d'une loi normale $\mathcal{N}(m, \sigma^2)$, il suffit d'estimer la moyenne et la variance des observations pour connaître la loi. En revanche, on prend le risque de considérer un modèle qui n'est pas exact. Si les observations ne sont pas Gaussiennes, on ne retrouvera jamais leur loi en se limitant à un modèle Gaussien.

Définition 1.2. *Un modèle $\mathcal{M}$ est un ensemble de lois possibles pour les observations. Le modèle est paramétrique s'il est indexé par un ensemble de dimension finie: $\mathcal{M} = \{\mu_\theta, \theta \in \Theta\}$, $\Theta \subseteq \mathbb{R}^d$. Dans le cas contraire, le modèle est non-paramétrique.*

Les familles de lois usuelles sont des exemples de modèles paramétriques (ex: modèle Gaussien $\mathcal{M} = \{\mathcal{N}(m, \sigma^2), m \in \mathbb{R}, \sigma^2 > 0\}$, modèle de Poisson $\mathcal{M} = \{\mathcal{P}(\lambda), \lambda > 0\}$, modèle exponentiel $\{\mathcal{E}(\lambda), \lambda > 0\}$ etc...). Les modèles non-paramétriques sont moins spécifiques et sont utilisés dans les situations où on dispose de peu d'information sur la loi μ des observations. Par exemple, μ est une loi continue (densité), μ admet un moment d'ordre deux, μ est une loi symétrique par rapport à zéro ou encore μ est une loi de probabilité (aucune information) sont des exemples de modélisation non-paramétriques.

En choisissant un modèle paramétrique, on prend le risque que la vraie loi des données n'appartienne pas au modèle. Un modèle non-paramétrique permet donc une évaluation parfois moins précise mais plus robuste de cette loi.

1.2 Paramètres et estimateurs

Définition 1.3. *Une statistique $T = T(X_1, ..., X_n)$ est une fonction (mesurable) connue des observations.*

Une statistique T est une variable aléatoire dont on observe une réalisation $t = T(x_1, ..., x_n)$.

Définition 1.4. *Un paramètre $\theta \in \Theta$ est une quantité inconnue qui dépend de la loi des observations.*

Formellement, un paramètre est l'image de μ par une fonctionnelle connue $\theta(.)$ (par exemple, l'espérance ou la médiane sous la loi μ). L'ensemble de définition du paramètre θ, noté Θ est donc l'image du modèle $\mathcal{M}$ par la fonctionnelle $\theta(.)$: $\Theta = \theta(\mathcal{M})$. Pour simplifier les notations, la dépendance en μ est souvent omise et θ désigne simplement une valeur de l'ensemble Θ, ou même la valeur associée à la vraie loi $\theta = \theta(\mu)$.

Une statistique utilisée pour estimer un paramètre θ est appelée *estimateur* et souvent notée $\hat{\theta}_n$. Un estimateur $\hat{\theta}_n$ d'un paramètre θ est dit *consistant* s'il converge en probabilité vers θ quand la taille de l'échantillon n tend vers l'infini, et ce quelque soit la vraie loi des observations. Si la convergence est presque sûre, alors $\hat{\theta}_n$ est dit fortement consistant.

S'il existe une suite réelles $(a_n)_{n\in\mathbb{N}}$ qui tend vers l'infini telle que la suite $a_n(\hat{\theta}_n - \theta)$ est bornée en probabilité, alors on dit que l'estimateur $\hat{\theta}_n$ converge vers θ à la vitesse a_n ou encore que $\hat{\theta}_n$ est a_n-consistant. La vitesse de convergence d'un estimateur $\hat{\theta}_n$ est une suite $(a_n)_{n\in\mathbb{N}}$ qui vérifie

$$a_n\big(\hat{\theta}_n - \theta\big) = O_p(1) \quad \text{et} \quad a_n\big(\hat{\theta}_n - \theta\big) \neq o_p(1).$$

Remarque 1.5. *Si un estimateur $\hat{\theta}_n$ est tel que $a_n\big(\hat{\theta}_n - \theta\big)$ converge en loi vers une limite non dégénérée (penser notamment au TCL avec la moyenne empirique qui converge vers l'espérance pour $a_n = \sqrt{n}$), alors $\hat{\theta}_n$ est a_n-consistant. La réciproque n'est évidemment pas vraie, une suite bornée en probabilité ne converge pas forcément en loi.*

1.3 Statistiques empiriques

Une idée standard de l'inférence statistique est de remplacer la loi théorique μ (loi des X_i) inconnue par la mesure empirique

$$\mu_n := \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i},$$

où δ est la masse de Dirac. C'est la loi uniforme sur l'échantillon lorsque celui-ci ne contient aucune valeur répétée, sinon le poids associé à une observation répétée k fois est k/n. La plupart des quantités relatives à la loi μ (paramètre) peuvent être estimées par leurs analogues empiriques, par exemple:

- La moyenne empirique $\hat{m} = \frac{1}{n}\sum_{i=1}^n X_i$ est un estimateur de l'espérance $m = \mathbb{E}_\mu(X)$ (quand elle existe). Il est sans biais et converge à vitesse $\sqrt{n}$ par le TCL si μ admet un moment d'ordre deux.
- La variance empirique $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \hat{m})^2$ est un estimateur consistant de la variance $\sigma^2 = \mathrm{var}_\mu(X)$ lorsque celle-ci est finie.
- La fonction de répartition empirique (dans le cas de X_i réels) $\hat{F} : t \mapsto = \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{X_i \leq t\}$ qui converge presque sûrement uniformément vers la fonction de répartition $F : t \mapsto \mathbb{P}(X \leq t)$ par le théorème de Glivenko-Cantelli.

Ces exemples sont fondés sur la même idée qui est de remplacer une quantité fonction de la loi théorique μ par la même fonction appliquée à μ_n. Par exemple, on vérifie bien que $\hat{m} = \int x d\mu_n(x)$, $\hat{F}(t) = \int_{-\infty}^t d\mu_n(s)$, etc...

Cette approche naturelle est très utilisée en pratique mais a ses limites. Par exemple, si μ admet une densité f (par rapport à la mesure de Lebesgue), l'approche empirique ne permet pas de l'estimer puisque μ_n n'a pas de densité.

Proposition 1.6. *Si les X_i sont iid de loi normale $\mathcal{N}(m, \sigma^2)$, les statistiques empiriques $\hat{m}$ et $\hat{\sigma}^2$ sont indépendantes.*

La preuve est admise (pour le moment).

Proposition 1.7. *Soit θ un paramètre de la forme $\theta = \int g(t)d\mu(t)$ où g est une fonction connue. Alors, la statistique empirique*

$$\hat{\theta} = \int g(s)d\mu_n(s) = \frac{1}{n}\sum_{i=1}^n g(X_i)$$

est un estimateur sans biais fortement consistant de θ.

Preuve. Conséquence directe de la loi forte des grands nombres. □

Exercice. Exprimer (si possible) les statistiques $\hat{m}$, $\hat{\sigma}^2$ et $\hat{F}(t)$ sous la forme $\int g(s)d\mu_n(s)$.

1.4 Statistiques d'ordre

Dans cette partie, on suppose les X_i à valeurs dans $\mathbb{R}$.

Définition 1.8. *Les statistiques d'ordre $X_{(1)}, ..., X_{(n)}$ sont les valeurs ordonnées dans l'ordre croissant de l'échantillon $X_1, ..., X_n$.*

Les statistiques d'ordre $X_{(k)}$ pour $k = 1, ..., n$ sont des variables aléatoires qui vérifient par construction $X_{(1)} \leq X_{(2)} \leq ... \leq X_{(n)}$ p.s.

Théorème 1.9. *Si $X_1, ..., X_n$ sont iid de densité f et fonction de répartition F, alors la k-ième statistique d'ordre $X_{(k)}$ a pour densité*

$$f_{(k)}(x) = \frac{n!}{(k-1)!(n-k)!} F(x)^{k-1}\big(1 - F(x)\big)^{n-k} f(x),$$

pour tout x point de continuité de f.

Preuve. Soit $h > 0$, on a pour tout $x \in \mathbb{R}$

$$\begin{aligned}
\mathbb{P}(x < X_{(k)} \leq x+h) &= \mathbb{P}\Big(\#\{i : X_i \leq x\} \leq k-1,\ \#\{i : X_i \leq x+h\} \geq k\Big) \\
&= \sum_{j=0}^{k-1} \sum_{\ell=k-j}^{n-j} \mathbb{P}\Big(\#\{i : X_i \leq x\} = j,\ \#\{i : x < X_i \leq x+h\} = \ell\Big) \\
&= \sum_{j=0}^{k-1} \sum_{\ell=k-j}^{n-j} \binom{n}{j}\binom{n-j}{\ell} F(x)^j \big(F(x+h) - F(x)\big)^\ell \big(1 - F(x)\big)^{n-j-\ell}.
\end{aligned}$$

On déduit

$$f_{(k)}(x) = \lim_{h \to 0} \frac{1}{h}\ \mathbb{P}(x < X_{(k)} \leq x+h) = \frac{n!}{(k-1)!(n-k)!} F(x)^{k-1}\big(1 - F(x)\big)^{n-k} f(x),$$

les termes pour $\ell \geq 2$ étant négligeables quand $h \to 0$. □

Exercice. Montrer que si les X_i sont iid de densité commune f strictement positive sur $\mathbb{R}$, alors pour tout $\alpha \in]0,1[$, $X_{\lceil n\alpha \rceil}$ converge vers le quantile d'ordre α quand $n \to \infty$, où $\lceil x \rceil$ est le plus petit entier supérieur ou égal à x.

Exercice. Proposer un estimateur consistant du paramètre de position θ dans le modèle de Cauchy avec X_i de densité f_θ donnée par

$$f_\theta(x) = \frac{1}{\pi} \frac{1}{1 + (\theta - x)^2}\ ,\ x \in \mathbb{R}.$$

2 Modèles paramétriques

2.1 Vraisemblance

Un modèle désigne une famille de lois associée aux observations X. Si cette famille est indexée par un paramètre $\theta \in \Theta$ de dimension finie où Θ est un sous-ensemble de $\mathbb{R}^p$, alors le modèle est dit paramétrique.

Définition 2.1. *Un modèle paramétrique $\mathcal{M} = \{\mu_\theta : \theta \in \Theta \subseteq \mathbb{R}^p\}$ est dit:*

- *identifié si l'application $\theta \mapsto \mu_\theta$ est injective sur Θ. Autrement dit, un modèle est identifié si chaque valeur du paramètre est associée à une unique loi du modèle.*
- *dominé si toutes les lois μ_θ du modèle sont absolument continues par rapport à une même mesure de référence ν.*

Dans le cas général μ_θ désigne la loi de X sous θ (donc une loi sur $\mathbb{R}^n$). Dans le contexte d'échantillonnage avec les X_i iid, la loi du vecteur X est entièrement caractérisée par la loi marginale des X_i. Dans ce cas, μ_θ fait référence à la loi marginale pour simplifier les notations.

Définition 2.2 (générale)**.** *Soit $\mathcal{M}$ un modèle paramétrique $\mathcal{M} = \{\mu_\theta : \theta \in \Theta\}$ identifié et dominé par une mesure ν. On note $p_\theta = d\mu_\theta/d\nu$ la densité de μ_θ par rapport ν. La vraisemblance du modèle est la fonction aléatoire*

$$\theta \mapsto p_\theta(X)\ ,\ \theta \in \Theta.$$

Dans le cadre d'échantillonnage avec les X_i iid et de loi continue ou à support dénombrable, la vraisemblance peut se définir plus simplement.

Définition 2.3 (simple)**.** *Soit un modèle d'échantillonnage avec $\mathcal{M} = \{\mu_\theta, \theta \in \Theta\}$ un modèle paramétrique identifié et $X = (X_1, ..., X_n)$ le vecteur des observations. On note f_θ:*

1. *la densité de μ_θ par rapport à la mesure de comptage dans le cas d'observations X_i discrètes, $f_\theta(x) = \mathbb{P}_\theta(X_i = x)$,*
2. *la densité de μ_θ par rapport à la mesure de Lebesgue dans le cas de variables X_i continues.*

La vraisemblance du modèle est la fonction

$$\theta \mapsto V(\theta, X) = \prod_{i=1}^{n} f_\theta(X_i), \ \theta \in \Theta.$$

Comme son nom l'indique, $V(\theta, X)$ donne la vraisemblance de θ étant donné X. En pratique on observe la réalisation $V(\theta, x)$ qui mesure la "probabilité" d'observer x si la vraie valeur du paramètre est θ (c'est rigoureusement la probabilité dans le cas discret avec ν la mesure de comptage et cette notion est généralisée à des densités par rapport à une mesure de référence quelconque).

On s'intéressera souvent à la *log-vraisemblance*

$$v(\theta, X) := \log\big(V(\theta, X)\big) = \sum_{i=1}^{n} \log\big(f_\theta(X_i)\big) \ , \ \theta \in \Theta$$

dont on étudiera les propriétés asymptotiques par la suite.

Il n'est utile de connaître la vraisemblance qu'à constante (positive) multiplicative près (remplacer la mesure de référence ν par 2ν revient à modifier la vraisemblance d'un facteur 2 sans en changer l'interprétation qu'on en fait). On peut le voir en regardant l'impact d'une transformation des variables sur la vraisemblance. Soit g un $\mathcal{C}^1$-diffeomorphisme qu'on applique aux données par $Y_i := g(X_i)$. Dans le cas d'un modèle dominé par la mesure de Lebesgue, la vraisemblance du vecteur $Y = (Y_1, ..., Y_n)$ est alors donnée par

$$\tilde{V}(\theta, Y) = \prod_{i=1}^{n} \frac{1}{|g' \circ g^{-1}(Y_i)|} f_\theta \circ g^{-1}(Y_i) = \frac{1}{|\prod_{i=1}^{n} g'(X_i)|} V(\theta, X).$$

Vues comme des fonctions de θ, les vraisemblances de Y et X sont proportionnelles. Or, la transformation des données n'a pas induit de perte d'information puisque g est bijective.

Une reparamétrisation du modèle n'impacte pas la vraisemblance lorsque celle-ci est bijective. Soit $\psi : \Theta \to \Psi$ une bijection, le modèle $\mathcal{M} = \{\mu_\theta : \theta \in \Theta\}$ paramétré par $\psi = \psi(\theta)$ s'écrit $\mathcal{M} = \{\mu_{\psi^{-1}(\psi)} : \psi \in \Psi\}$ et la vraisemblance $V_\psi(., X) = V(\psi^{-1}(.), X)$ (penser par exemple au modèle Gaussien de moyenne connue qui peut être paramétré par la variance ou l'écart-type). En revanche, une transformation du paramètre peut rendre un modèle paramétrique non identifié si celle-ci n'est pas injective. C'est évidemment à éviter, mais dans ce cas, il est possible de définir la vraisemblance par

$$V_\psi(\psi, X) = \sup_{\theta : \psi(\theta) = \eta} V(\theta, X).$$

2.2 Familles exponentielles

Définition 2.4. *Un modèle paramétrique de vraisemblance $V(.,X)$ est une famille exponentielle s'il existe des fonctions $G : \mathbb{R}^n \to \mathbb{R}_+$, $H : \Theta \to \mathbb{R}_+$, $T : \mathbb{R}^n \to \mathbb{R}^q$ et $\eta : \Theta \to \mathbb{R}^q$ telle que*

$$V(\theta, X) = G(X)H(\theta) \exp\big(\langle T(X), \eta(\theta)\rangle\big).$$

On remarque que cette écriture n'est pas unique.

Dans une famille exponentielle, la statistique $T(x) = (T_1(X), ..., T_q(X))$ est appelée statistique naturelle et $\eta(\theta) = (\eta_1(\theta), ..., \eta_q(\theta))$ le paramètre naturel associé. Ce sont les termes les plus importants dans l'écriture d'une famille exponentielle car ils résument en un sens toute l'information nécessaire à l'inférence, comme nous le verrons par la suite. La taille q des vecteurs $T(X)$ et $\eta(\theta)$ est la *dimension* du modèle qui est le plus souvent égale à la dimension p du modèle de départ où $\Theta \subseteq \mathbb{R}^p$.

La statistique $G(X)$ a en revanche peu d'intérêt, elle peut être traitée comme une constante puisque la vraisemblance est une fonction de θ. Enfin, le terme $H(\theta)$ peut être vu comme une constante de normalisation qui a aussi peu d'intérêt en pratique.

Dans le modèle d'échantillonnage iid, la décomposition se fait sur la densité marginale f_θ (par rapport à la mesure de référence ν):

$$f_\theta(x) = g(x)h(\theta) \exp\big(\langle t(x), \eta(\theta)\rangle\big),$$

On peut alors identifier $G(X) = \prod_{i=1}^n g(X_i)$, $H(\theta) = h(\theta)^n$ et $T(X) = \sum_{i=1}^n t(X_i)$, le paramètre naturel $\eta(\theta)$ est le même dans les deux cas.

Définition 2.5. *Une famille exponentielle est dite:*

- *canonique si θ est le paramètre naturel (la fonction $\eta(.)$ est l'identité),*
- *standard si $T(X) = \sum_{i=1}^n X_i \in \mathbb{R}^k$ est la statistique naturelle.*

Toute famille exponentielle peut être rendue canonique en reparamétrant le modèle par $\eta = \eta(\theta)$, on parle alors de représentation canonique. On peut de plus se ramener au cas $G(X) = 1$ en modifiant la mesure de référence ν par $G\nu$, ce qui ne change pas fondamentalement la vraisemblance. On peut alors écrire

$$f_\eta(x) = \tilde{h}(\eta) \exp\big(\langle T(x), \eta\rangle - \zeta(\eta)\big),$$

où $\zeta(\eta) = \ln\big(\int \exp(\langle T(x), \eta\rangle) g(x) d\nu(x)\big)$.

Exercice 1. Montrer que les modèles suivants sont des familles exponentielles:

1. Modèle Gaussien $\mathcal{M} = \{\mathcal{N}(m, \sigma^2) : m \in \mathbb{R}, \sigma^2 > 0\}$.
2. Modèle de Poisson $\mathcal{M} = \{\mathcal{P}(\lambda) : \lambda > 0\}$.
3. Modèles binomiaux $\mathcal{M} = \{\mathcal{B}(n, p) : p \in]0, 1[\}$ avec n fixé.

Exercice 2. Montrer qu'un modèle paramétrique de densité f_θ dont le support dépend de θ ne peut pas être une famille exponentielle.

2.3 Vecteurs Gaussiens

Sous sa forme vectorielle, la limite est une généralisation vectorielle de la loi normale, appelée vecteur Gaussien. Par convention, on considèrera une constante comme une variable aléatoire Gaussienne de variance nulle.

Définition 2.6. *Soit $m \in \mathbb{R}^k$ et $\Sigma \in \mathbb{R}^{k\times k}$ une matrice symmétrique semi définie-positive. Un vecteur aléatoire X de $\mathbb{R}^k$ est un vecteur Gaussien de moyenne m et de variance Σ si pour tout vecteur $a \in \mathbb{R}^k$, $a^\top X \sim \mathcal{N}(a^\top m, a^\top \Sigma a)$. On le note $X \sim \mathcal{N}(m, \Sigma)$.*

On vérifie facilement que la matrice Σ est bien la matrice de covariance de X. En effet,

- Pour $a = e_i$ le i-ème vecteur de la base canonique, $a^\top X = X_i \sim \mathcal{N}(m_i, \Sigma_{ii}) \Longrightarrow \mathrm{var}(X_i) = \Sigma_{ii}$.
- Pour $a = e_i + e_j$, on obtient $X_i + X_j \sim \mathcal{N}(m_i + m_j, \Sigma_{ii} + \Sigma_{jj} + 2\Sigma_{ij})$ d'où on déduit $\mathrm{cov}(X_i, X_j) = \Sigma_{ij}$.

Une propriété fondamentale de la famille des vecteurs Gaussiens est qu'elle est stable par transformation affine.

Proposition 2.7. *Soit $X \sim \mathcal{N}(m, \Sigma)$ un vecteur Gaussien de $\mathbb{R}^k$. Pour $A \in \mathbb{R}^{p\times k}$ et $b \in \mathbb{R}^p$, $Y = AX + b$ est un vecteur Gaussien de $\mathbb{R}^p$ de loi $Y \sim \mathcal{N}(Am + b, A\Sigma A^\top)$.*

Preuve. Soit $a \in \mathbb{R}^p$, on a $a^\top Y = a^\top(AX + b) = (A^\top a)^\top X + a^\top b$. Comme $X \sim \mathcal{N}(m, \Sigma)$, on sait que $(A^\top a)^\top X \sim \mathcal{N}\big((A^\top a)^\top m, (A^\top a)^\top \Sigma (A^\top a)\big)$. Comme $a^\top b$ est une constante, on déduit

$$(A^\top a)^\top X + a^\top b = AY + b \sim \mathcal{N}\big(a^\top(Am + b), a^\top(A\Sigma A^\top)a\big).$$

Ceci est valable pour tout $a \in \mathbb{R}^p$, d'où le résultat. □

Il existe d'autres définitions équivalentes d'un vecteur Gaussien.

Proposition 2.8. *Un vecteur aléatoire X est un vecteur Gaussien de moyenne $m \in \mathbb{R}^k$ et de variance $\Sigma \in \mathbb{R}^{k\times k}$ si, et seulement si:*

1. *X a pour fonction caractéristique*

$$\varphi_X(t) = \mathbb{E}\big(e^{it^\top X}\big) = \exp\Big(it^\top m - \frac{t^\top \Sigma t}{2}\Big)\ ,\ t \in \mathbb{R}^k.$$

2. *Si Σ est inversible, X a pour densité*

$$f_X(x) = \frac{1}{(2\pi)^{k/2}\sqrt{\det(\Sigma)}} \exp\Big(-\frac{1}{2}(x-m)^\top \Sigma^{-1}(x-m)\Big)\ ,\ x \in \mathbb{R}^k.$$

Idée de preuve. Soit X qui vérifie la définition 2.6, on remarque que

$$\varphi_X(t) = \varphi_{t^\top X}(1) = \exp\Big(i\ t^\top m - \frac{t^\top \Sigma t}{2}\Big),$$

où le terme de droite est une simple évaluation de la fonction caractéristique d'une variable aléatoire Gaussienne réelle $\mathcal{N}(t^\top m, t^\top \Sigma t)$. □

Remarque 2.9. *Lorsque la matrice Σ n'est pas inversible, X "vit" dans le sous-espace affine $\mathcal{A} = \{m\} \oplus \mathrm{Im}(\Sigma) \subset \mathbb{R}^k$, dans le sens où $\mathbb{P}(X \in \mathcal{A}) = 1$. Dans ce cas, X n'admet pas de densité sur $\mathbb{R}^k$ par rapport à la mesure de Lebesgue.*

Proposition 2.10. *Soit $X \sim \mathcal{N}(m, \Sigma)$ et $Y_1 = A_1 X + b_1$, $Y_2 = A_2 X + b_2$ deux transformations affines de X à valeurs dans $\mathbb{R}^{p_1}$ et $\mathbb{R}^{p_2}$ respectivement. Les énoncés suivants sont équivalents:*

i) Y_1 *et* Y_2 *sont indépendants*

ii) Y_1 *et* Y_2 *sont non-corrélés:* $\mathbb{E}(Y_1 Y_2^\top) - \mathbb{E}(Y_1)\mathbb{E}(Y_2^\top) = 0 \in \mathbb{R}^{p_1 \times p_2}$

iii) $A_1 \Sigma A_2^\top = 0 \in \mathbb{R}^{p_1 \times p_2}$

En particulier, les coordonnées d'un vecteur Gaussien sont indépendantes si et seulement si elles sont non-corrélées.

Preuve. Clairement $i) \Longrightarrow ii)$. De plus, $\mathbb{E}(Y_1 Y_2^\top) - \mathbb{E}(Y_1)\mathbb{E}(Y_2^\top) = A_1 \Sigma A_2^\top$ d'où $ii) \Longrightarrow iii)$. Enfin, $Y := (Y_1^\top, Y_2^\top)^\top$ est un vecteur Gaussien par la proposition 2.7. Pour $t = (t_1, t_2)^\top \in \mathbb{R}^{p_1+p_2}$, sa fonction caractéristique est donnée par

$$\varphi_Y(t) = \exp\Big(i\ (t_1^\top b_1 + t_2^\top b_2) - \frac{t_1^\top A_1 \Sigma A_1^\top t_1 + t_2^\top A_1 \Sigma A_1^\top t_1 + 2 t_1^\top A_1 \Sigma A_2^\top t_2}{2}\Big).$$

On a donc $\varphi_Y(t) = \varphi_{Y_1}(t_1)\varphi_{Y_2}(t_2)$ si et seulement si $A_1 \Sigma A_2^\top = 0$. □

Théorème 2.11 (Théorème central limite)**.** *Soient $X_1, ..., X_n, ...$ des variables aléatoires de $\mathbb{R}^k$ iid de carré intégrable telles que $\mathbb{E}(X_i) = m \in \mathbb{R}^k$ et $\mathrm{var}(X_i) = \Sigma \in \mathbb{R}^{k \times k}$, l'estimateur $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ vérifie*

$$\sqrt{n}(\overline{X}_n - m) \xrightarrow[n\to\infty]{loi} \mathcal{N}(0, \Sigma).$$

Une preuve simple du théorème central limite utilise le théorème de Lévy, en développant à l'ordre 2 la fonction caractéristique au voisinage de zéro.

Preuve. On pose $Z_i = X_i - m$. On a $\mathbb{E}(Z_i) = 0$, $\mathbb{E}(Z_i Z_i^\top) = \Sigma$ et

$$\varphi_Z(t) = \mathbb{E}(e^{i\langle t, Z_1\rangle}) = 1 - \frac{t^\top \Sigma t}{2} + o(\|t\|^2).$$

En écrivant $\sqrt{n}(\overline{X}_n - m) = \frac{1}{\sqrt{n}}\sum_{i=1}^n Z_i$, on déduit

$$\varphi_{\sqrt{n}(\overline{X}_n - m)}(t) = \mathbb{E}\big(e^{i\langle t, \sqrt{n}(\overline{X}_n - m)\rangle}\big) = \prod_{i=1}^n \mathbb{E}\big(e^{i\frac{1}{\sqrt{n}}\langle t, Z_i\rangle}\big) = \Big[\varphi_Z\Big(\frac{t}{\sqrt{n}}\Big)\Big]^n = \Big[1 - \frac{t^\top \Sigma t}{2n} + o\Big(\frac{\|t\|^2}{n}\Big)\Big]^n.$$

Pour tout $t \in \mathbb{R}^k$, la fonction caractéristique de $\sqrt{n}(\overline{X}_n - m)$ au point t converge donc vers $e^{-\frac{1}{2}t^\top \Sigma t}$, la fonction caractéristique d'une loi normale $\mathcal{N}(0, \Sigma)$. On conclut par le théorème de Lévy. □

3 Information et exhaustivité

3.1 Information de Fisher

L'information de Fisher d'un modèle paramétrique $\mathcal{M} = \{\mu_\theta : \theta \in \Theta\}$ permet de mesurer la quantité d'information du modèle via sa vraisemblance. C'est une notion locale qui vise à évaluer le pouvoir de discrimination du modèle entre deux valeurs proche du paramètre θ. Elle nécessite des conditions du régularité sur le modèle, notamment de dérivabilité de la log-vraisemblance par rapport à θ.

L'information de Fisher est construite à partir de la vraisemblance, elle adhère donc au postulat que toute l'information des données dans un modèle paramétrique est contenue dans sa vraisemblance (parfois appelé principe de vraisemblance).

Définition 3.1. *Soit $\mathcal{M} = \{\mu_\theta : \theta \in \Theta \subseteq \mathbb{R}^p\}$ un modèle paramétrique dominé et $v(\theta, X)$ sa log-vraisemblance. On suppose que $v(., x)$ est différentiable pour ν-presque tout x. On appelle score du modèle au point θ le gradient (par rapport à θ) de la log-vraisemblance $S_\theta(X) = \nabla v(\theta, X)$, qui est à valeurs dans $\mathbb{R}^p$.*

On peut montrer que le score est identique quelque soit la mesure dominante ν choisie (autrement dit deux vraisemblances proportionnelles ont le même score), il s'agit donc d'une notion intrinsèque au modèle.

Proposition 3.2. *Sous des conditions de dérivabilité sous le signe intégral de la vraisemblance:*

$$\forall \theta \in \Theta, \int \nabla V(\theta, x) d\nu(x) = \nabla \int V(\theta, x) d\nu(x),$$

le score est d'espérance nulle sous θ: $\mathbb{E}_\theta(S_\theta(X)) = 0$.

Preuve. Il suffit d'écrire

$$\mathbb{E}_\theta(S_\theta(X)) = \int \frac{\nabla V(\theta, x)}{V(\theta, x)} V(\theta, x) d\nu(x) = \int \nabla V(\theta, x) d\nu(x) = \nabla \int V(\theta, x) d\nu(x).$$

On conclut en remarquant que $\int V(\theta, x) d\nu(x) = 1$ pour tout θ. □

Pour pouvoir dériver sous le signe intégral en θ, il suffit ici que les dérivées partielles de la vraisemblance soient dominées par une fonction h intégrable dans un voisinage $\mathcal{V}$ de θ. En notant ∂_{θ_i} l'opérateur de dérivée partielle par rapport à θ_i,

$$\forall \theta \in \mathcal{V} \ , \ \left|\partial \theta_i V(\theta, x)\right| \leq h(x) \ \text{ avec } \ \int h(x) d\nu(x) < \infty.$$

Ces hypothèses techniques de dérivabilité contribuent à la régularité du modèle. Précisément, un modèle est dit *régulier* si le score existe, Θ est un sous-ensemble ouvert de $\mathbb{R}^p$ et si pour toute fonction h μ_θ-intégrable pour tout $\theta \in \Theta$,

1. $\partial_{\theta_i} \int h(x) V(\theta, x) d\nu(x) = \int h(x) \ \partial_{\theta_i} V(\theta, x) d\nu(x) \ , \ i = 1, ..., p$
2. $\partial_{\theta_i} \partial_{\theta_j} \int h(x) V(\theta, x) d\nu(x) = \int h(x) \ \partial_{\theta_i} \partial_{\theta_j} V(\theta, x) d\nu(x) \ , \ i, j = 1, ..., p.$

On remarque que $\int h(x)V(\theta,x)d\nu(x) = \mathbb{E}_\theta(h(X))$.

Définition 3.3. *Soit $\mathcal{M} = \{\mu_\theta : \theta \in \Theta \subseteq \mathbb{R}^p\}$ un modèle paramétrique de fonction de score $S_\theta(X)$ que l'on suppose de carré intégrable sous μ_θ pour tout $\theta \in \Theta$. L'information de Fisher $\mathcal{I}$ est la fonction qui à θ associe*

$$\mathcal{I}(\theta) := \mathrm{var}_\theta\left(S_\theta(X)\right).$$

L'information de Fisher est donc une matrice semi définie-positive de taille $p \times p$, où p est la dimension du paramètre θ.

Proposition 3.4. *Dans $\mathcal{M} = \{\mu_\theta : \theta \in \Theta \subseteq \mathbb{R}^p\}$ un modèle régulier et $H_\theta(X)$ la matrice Hessienne de $v(\theta, X)$ par rapport à θ,*

$$\mathcal{I}(\theta) = -\mathbb{E}_\theta\big(H_\theta(X)\big).$$

Preuve. Par la proposition 3.2, $S_\theta(X)$ est sans biais sous μ_θ d'où

$$\mathcal{I}(\theta)_{ij} = \mathrm{var}_\theta\left(S_\theta(X)\right)_{ij} = \mathbb{E}_\theta\big(\partial_{\theta_i} v(\theta,X)\partial_{\theta_j} v(\theta,X)\big) = \int \frac{\partial_{\theta_i} V(\theta,x)\ \partial_{\theta_j} V(\theta,x)}{V(\theta,x)} d\nu(x).$$

D'autre part,

$$H_\theta(X)_{ij} = \partial_{\theta_i}\partial_{\theta_j} v(\theta,X) = \partial_{\theta_i}\frac{\partial_{\theta_j} V(\theta,X)}{V(\theta,X)} = \frac{\partial_{\theta_i}\partial_{\theta_j} V(\theta,X)}{V(\theta,X)} - \frac{\partial_{\theta_i} V(\theta,X)\ \partial_{\theta_j} V(\theta,X)}{V(\theta,X)^2}.$$

En intégrant contre $\mu_\theta = V(\theta,.)\nu$,

$$\mathbb{E}_\theta\big(H_\theta(X)_{ij}\big) = \int \partial_{\theta_i}\partial_{\theta_j} V(\theta,x)d\nu(x) - \int \frac{\partial_{\theta_i} V(\theta,x)\ \partial_{\theta_j} V(\theta,x)}{V(\theta,x)} d\nu(x).$$

Comme le modèle est régulier, $\int \partial_{\theta_i}\partial_{\theta_j} V(\theta,x)d\nu(x) = \partial_{\theta_i}\partial_{\theta_j}\int V(\theta,x)d\nu(x) = 0$. □

En pratique, l'information de Fisher peut donc se calculer de deux façons différentes dans un modèle régulier.

Proposition 3.5. *Soit $X_1, ..., X_n$ un échantillon iid dans un modèle paramétrique régulier et $\mathcal{I}_k(\theta)$ l'information de Fisher pour les k premières observations. Alors*

$$\mathcal{I}_k(\theta) = k\mathcal{I}_1(\theta).$$

Preuve. Exercice. □

3.2 Conditionnement et Exhaustivité

Si on postule que l'information contenue dans la vraisemblance ne dépend pas de la mesure de référence ν, deux vraisemblances proportionnelles (vues comme des fonctions de θ) contiennent la même information. En particulier, appliquer une transformation aux données par $Y = g(X)$ n'induit pas de perte d'information sur θ si la vraisemblance de Y est proportionnelle à celle de X. La notion d'exhaustivité, définie à partir des lois conditionnelles (que l'on définira dans les cas discret et continu uniquement) est basée sur ce principe.

Notation. Si ν est la mesure de comptage, $f_X(x) = \mathbb{P}(X = x)$ désigne la densité par rapport à ν d'une variable X à valeurs dans un espace dénombrable $\mathcal{X}$. Pour toute fonction h f_X-intégrable, on utilise la notation

$$\mathbb{E}(h(X)) = \int h(x) f_X(x) d\nu(x) = \sum_{x \in \mathcal{X}} h(x) \mathbb{P}(X = x).$$

Définition 3.6. *Soit (X, Y) un couple de variables aléatoires de densités marginales f_X, f_Y par rapport à ν_X, ν_Y respectivement et de densité jointe $f_{X,Y}$ par rapport à $\nu_X \otimes \nu_Y$. La loi conditionnelle de Y sachant X a pour densité par rapport à ν_Y*

$$f_{Y|X} : y \mapsto \frac{f_{X,Y}(X, y)}{f_X(X)}.$$

La loi conditionnelle de Y sachant X est une loi aléatoire puisque fonction X. Dans le cas discret, elle s'obtient à partir de la formule de Bayes en considérant la probabilité conditionnelle

$$(x, y) \mapsto \mathbb{P}(Y = y | X = x) = \frac{\mathbb{P}(Y = y, X = x)}{\mathbb{P}(X = x)}$$

appliquée en $x = X$.

Exercice. Calculer la loi conditionnelle de Y sachant X dans les cas suivants:

1. Y et Z sont indépendantes, de loi de Poisson de paramètre 1 et $X = Y + Z$.
2. (X, Y) est un vecteur Gaussien centré tel que $\mathrm{var}(X) = \mathrm{var}(Y) = 1$ et $\mathrm{cov}(X, Y) = -1/2$.
3. Z est de loi exponentielle de paramètre 1, $X = \lfloor Z \rfloor$ (la partie entière de Z) et $Y = Z - X$.

On définit maintenant l'espérance conditionnelle de Y sachant X que l'on note $\mathbb{E}(Y|X)$. C'est une variable aléatoire qui peut s'écrire comme une fonction mesurable de X. Il existe trois caractérisations importantes.

i) Si la loi conditionnelle de Y sachant X admet un moment d'ordre 1 presque-sûrement, alors l'espérance conditionnelle de Y sachant X est définie comme l'espérance sous la loi conditionnelle, typiquement

$$\mathbb{E}(Y|X) = \int y f_{Y|X}(y) d\nu_Y(y).$$

ii) Si Y est intégrable, alors $\mathbb{E}(Y|X)$ est l'unique (p.s.) variable aléatoire fonction de X qui vérifie pour toute fonction h telle que $\mathbb{E}|h(X)Y| < \infty$,

$$\mathbb{E}\big(h(X)Y\big) = \mathbb{E}\big(h(X)\mathbb{E}(Y|X)\big).$$

iii) Si Y est de carré intégrable, alors $\mathbb{E}(Y|X) = \rho^*(X)$ où ρ^* minimise $\rho \mapsto \mathbb{E}(Y - \rho(X))^2$ parmi les fonctions mesurables de X.

Proposition 3.7. *L'espérance conditionnelle vérifie, sous réserve d'existence:*

- $\mathbb{E}\big(\mathbb{E}(Y|X)\big) = \mathbb{E}(Y)$.

- *Pour tout* $\lambda \in \mathbb{R}$, $\mathbb{E}(Y+\lambda Z|X) = \mathbb{E}(Y|X)+\lambda\mathbb{E}(Z|X)$ *(linéarité de l'espérance conditionnelle).*
- *Pour toute fonction* g *telle que* $\mathbb{E}|Yg(X)| < \infty$, $\mathbb{E}(Yg(X)|X) = g(X)\mathbb{E}(Y|X)$.
- X *et* Y *sont indépendantes si et seulement si* $\mathbb{E}(h(Y)|X) = \mathbb{E}(h(Y))$ *pour toute fonction* h *telle que* $\mathbb{E}|h(Y)| < \infty$.
- Y *est presque sûrement une fonction de* X *si et seulement si* $\mathbb{E}(Y|X) \overset{ps}{=} Y$.

Preuve. Exercice. □

On n'a défini la notion de loi conditionnelle que dans les cas continus et discret mais celle-ci peut être définie dans le cas général en utilisant l'espérance conditionnelle. Il suffit de poser pour tout ensemble mesurable A,

$$\mathbb{P}(Y \in A|X) = \mathbb{E}\big(\mathbb{1}\{Y \in A\}\big|X\big).$$

Définition 3.8. *Soit* $\mathcal{M} = \{\mu_\theta : \theta \in \Theta \subseteq \mathbb{R}^p\}$ *un modèle paramétrique, une statistique* $S(X)$ *est exhaustive si la loi de* X *sachant* $S(X)$ *ne dépend pas de* θ.

Dans le cas discret, la loi conditionnelle est obtenue par la formule de Bayes

$$P_\theta(X = x|S(X) = s) = \frac{\mathbb{P}_\theta(X = x, S(X) = s)}{\mathbb{P}_\theta(S(X) = s)} = \frac{\mathbb{P}_\theta(X = x)\mathbb{1}\{S(x) = s\}}{\mathbb{P}_\theta(S(X) = s)}.$$

La statistique $S(X)$ est donc exhaustive si pour tout x tel que $S(x) = s$, $\mathbb{P}_\theta(X = x)$ est proportionnel à $\mathbb{P}_\theta(S(X) = s)$ (vue comme des fonction de θ, c'est-à-dire que le ratio ne dépend pas de θ). Dans le cas général, calculer la loi conditionnelle de X sachant $S(X)$ est plus difficile. Par exemple, pour des variables continues, la définition précédente de s'applique pas car $(X, S(X))$ n'admet pas de densité par rapport à la mesure de Lebesgue sur $\mathcal{X} \times S(\mathcal{X})$. Mais, l'argument de proportionnalité reste valable: si $V_S(\theta, S(X))$ est la vraisemblance de $S(X)$, la statistique $S(X)$ est exhaustive si $V(\theta, X)/V_S(\theta, S(X))$ ne dépend pas de θ. Autrement dit, que ce soit dans le cas continu ou discret, $S(X)$ est une statistique exhaustive si la vraisemblance de $S(X)$ est proportionnelle à celle de X et donc si le changement de variable $Y = S(X)$ n'induit pas de perte d'information.

Exercice. Montrer que les statistiques suivantes sont exhaustives:

1. Le nombre de succès dans un modèle d'échantillonnage de Bernoulli.
2. La moyenne empirique dans le modèle d'échantillonnage de Poisson.
3. La moyenne empirique dans le modèle d'échantillonnage Gaussien avec variance connue.

Théorème 3.9. *Une statistique* $S(X)$ *est exhaustive si et seulement si il existe des fonctions* h *et* g *telles que*

$$V(\theta, X) = h(X)g(\theta, S(X)).$$

Autrement dit, une statistique est exhaustive si la vraisemblance peut s'écrire comme une fonction de celle-ci et de θ, à une constante multiplicative près. Ce résultat, que l'on admettra, est appelé critère de factorisation ou critère de Fisher-Neyman. Cela implique en particulier que toute transformation bijective (mesurable) de X est une statistique exhaustive.

Exercice. Montrer que dans un modèle d'échantillonnage iid, le vecteur des statistiques d'ordre $(X_{(1)}, ..., X_{(n)})$ est une statistique exhaustive.

Corollaire 3.10. *Dans une famille exponentielle, la statistique naturelle est exhaustive.*

Preuve. Immédiat. □

L'intérêt d'une statistique exhaustive réside dans son potentiel de réduction de données. Par exemple, dans un modèle d'échantillonnage de Bernoulli $X_1, ..., X_n$ iid de loi $\mathcal{B}(p)$, on sait que toute l'information sur p est contenue dans le nombre de succès (autrement dit $S(X) = \sum_{i=1}^n X_i$ est exhaustive). On peut donc passer de n variables de Bernoulli à une variable binomiale sans perte d'information. La minimalité d'une statistique exhaustive est une notion d'optimalité de cette réduction de données.

Définition 3.11. *Une statistique exhaustive $S(X)$ est minimale si pour toute statistique exhaustive $S'(X)$, il existe une fonction ϕ telle que $S(X) = \phi(S'(X))$ μ_θ-presque-sûrement pour tout θ.*

Une statistique exhaustive est donc minimale s'il n'existe pas de transformation $\psi(S(X))$ non-réversible qui préserve l'exhaustivité. Une telle statistique n'est bien sûr pas unique, toute transformation bijective de celle-ci est également exhaustive minimale.

Théorème 3.12. *Soit S une fonction mesurable de l'échantillon, si on a l'équivalence*

$$S(x) = S(y) \iff \frac{V(\theta, x)}{V(\theta, y)} \text{ ne dépend pas de } \theta, \tag{$*$}$$

alors $S(X)$ est une statistique exhaustive minimale pour θ.

Preuve. Supposons $(*)$. Pour tout s dans l'image de $S(.)$, soit x_s tel que $S(x_s) = s$. On a

$$V(\theta, X) = \frac{V(\theta, X)}{V(\theta, x_{S(X)})} \times V(\theta, x_{S(X)}),$$

où $V(\theta, X)/V(\theta, x_{S(X)})$ ne dépend pas de θ et $V(\theta, x_{S(X)})$ est fonction de θ et $S(X)$. Par le critère de factorisation, $S(X)$ est donc exhaustive. Soit $S'(X)$ une statistique exhaustive et h', g' tels que

$$V(\theta, x) = h'(x)g'(\theta, S'(x)).$$

Pour tous x, y tels que $S'(x) = S'(y)$, on a

$$\frac{V(\theta, x)}{V(\theta, y)} = \frac{h'(x)g'(\theta, S'(x))}{h'(y)g'(\theta, S'(y))} = \frac{h'(x)}{h'(y)}$$

qui ne dépend donc pas de θ. On déduit par $(*)$ que $S(x) = S(y)$ et donc que S peut s'écrire comme une fonction de S'. □

Exercice. Calculer le rapport $V(m, \sigma^2, X)/V(m, \sigma^2, Y)$ dans le modèle Gaussien $\mathcal{N}(m, \sigma^2)$ avec m et σ^2 inconnus, et déduire que la statistique $S(X) = \left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2\right)$ est exhaustive minimale.

Exercice. Soient des X_i iid de densité $f_\theta(x) = (\ln(2)x)^{-1}\mathbb{1}\{x \in]\theta, 2\theta[\}$, montrer que la statistique $S(X) = (X_{(1)}, X_{(n)})$ est exhaustive minimale.

Dans ce dernier exemple, on remarque que la dimension d'une statistique exhaustive minimale ne coïncide pas toujours avec celle du paramètre.

Définition 3.13. *Dans un modèle paramétrique, une statistique $S(X)$ est complète si pour toute fonction ϕ intégrable*

$$\forall \theta \in \Theta \ , \ \mathbb{E}_\theta\big(\phi(S(X)\big) = 0 \implies \forall \theta \in \Theta \ , \ \mathbb{P}_\theta\big(\phi(S(X)) = 0\big) = 1.$$

Intuitivement, une statistique $S(X)$ est complète si toute transformation non-constante de $S(X)$ conserve de l'information sur θ.

Exercice. Dans le modèle d'échantillonnage de Bernoulli, montrer que $S(X) = \sum_{i=1}^n X_i$ est complète.

Proposition 3.14. *Dans une famille exponentielle $f_\theta(x) = g(x)h(\theta)\exp(\langle T(x), \eta(\theta)\rangle)$, si $\eta(\Theta) \subseteq \mathbb{R}^p$ est d'intérieur non-vide, alors la statistique naturelle $T(X)$ est complète.*

Preuve. Admise.

Définition 3.15. *Une statistique est libre si sa loi est constante sous toutes les lois du modèle.*

Une statistique libre est en quelque sorte l'opposé d'une statistique exhaustive dans le sens où elle n'apporte aucune information sur le paramètre du modèle.

Exercice. Trouver une statistique libre (non triviale) dans les modèles suivants:

1. Le modèle Gaussien sur la moyenne $\mathcal{M} = \{\mathcal{N}(\theta, 1) : \theta \in \mathbb{R}\}$.
2. Le modèle uniforme sur $[0, \theta], \theta > 0$.

Théorème 3.16 (Basu)**.** *Dans un modèle paramétrique $\mathcal{M} = \{\mu_\theta : \theta \in \Theta\}$, si $S(X)$ est une statistique exhaustive complète et $S'(X)$ une statistique libre, alors $S(X)$ et $S'(X)$ sont indépendantes sous μ_θ pour tout $\theta \in \Theta$.*

Preuve. Pour toute fonction h telle que $\mathbb{E}(h(S'))$ existe, soit ϕ_h telle que

$$\mathbb{E}_\theta\big(h(S')|S\big) - \mathbb{E}_\theta\big(h(S')\big) = \phi_h(S).$$

On vérifie bien que cette fonction ne dépend pas de θ du fait que S est exhaustive et S' est libre. On a

$$\mathbb{E}_\theta\big(\phi_h(S)\big) = \mathbb{E}_\theta\big(\mathbb{E}_\theta(h(S')|S) - \mathbb{E}_\theta(h(S'))\big) = \mathbb{E}_\theta(h(S')) - \mathbb{E}_\theta(h(S')) = 0.$$

La statistique S étant complète, on déduit que $\phi_h(S) = 0$ μ_θ-presque sûrement pour tout θ, et donc que $\mathbb{E}_\theta\big(h(S')|S\big) = \mathbb{E}_\theta\big(h(S')\big)$. Comme le résultat est vrai pour toute fonction h, on conclut que S' et S sont indépendantes pour tout θ. □

Exercice. Montrer que les moyenne et variance empririques sont indépendantes dans le modèle Gaussien (indication: se placer dans le modèle Gaussien avec variance connue $\mathcal{M} = \{\mathcal{N}(\theta, \sigma^2) : \theta \in \mathbb{R}\}$).

3.3 Efficacité

Le biais d'un estimateur $\hat{\psi}$ d'un paramètre $\psi = \psi(\theta)$ est la quantité $b_\theta(\hat{\psi}) = \mathbb{E}_\theta(\hat{\psi}) - \psi$ qui n'existe que si l'espérance de $\hat{\psi}$ est définie, c'est-à-dire si $\mathbb{E}_\theta|\hat{\psi}| < \infty$. Si le biais est nul pour tout θ, on dira

que $\hat{\psi}$ est un estimateur sans biais.

L'erreur quadratique moyenne (EQM) d'un estimateur $\hat{\psi}$ de ψ est la quantité

$$\mathrm{EQM}_\theta(\hat{\psi}) = \mathbb{E}_\theta(\hat{\psi} - \psi)^2 \in [0, +\infty].$$

On dit que $\hat{\psi}$ converge dans $\mathbb{L}^2$ si son EQM tend vers zéro (pour tout $\mu_\theta \in \mathcal{M}$).

Lemme 3.17 (Décomposition biais-variance)**.** *Pour tout estimateur $\hat{\psi}$ de carré intégrable,*

$$\mathrm{EQM}_\theta(\hat{\psi}) = b_\theta(\hat{\psi})^2 + \mathrm{var}_\theta(\hat{\psi}).$$

La formule de décomposition biais-variance permet de montrer immédiatement qu'un estimateur converge dans $\mathbb{L}^2$ si et seulement si son biais et sa variance tendent simultanément vers zéro. Comme la convergence dans $\mathbb{L}^2$ implique la convergence en probabilité, un estimateur qui converge en moyenne quadratique est consistant.

Dans un modèle, la question du meilleur estimateur au sens de l'EQM n'a (en général) pas de sens puisque la qualité d'un estimateur dépend de la vraie valeur du paramètre. Par exemple, si $\Theta = \mathbb{R}$, l'estimateur $\hat{\theta} = 3$ est le meilleur possible si 3 est la vraie valeur du paramètre (son EQM vaut 0) mais pas pour d'autres valeurs. Il n'existe pas d'estimateur optimal $\hat{\theta}$ qui minimiserait l'EQM pour toute valeur de θ. Les choses sont différentes si on se limite aux estimateurs sans biais.

Théorème 3.18 (Inégalité de Cramer-Rao)**.** *Soit $\mathcal{M} = \{\mu_\theta : \theta \in \Theta \subseteq \mathbb{R}\}$ un modèle régulier d'information de Fisher $\mathcal{I}(\theta)$ et $\hat{\theta} = \hat{\theta}(X)$ un estimateur <u>sans biais</u> de θ de carré intégrable pour tout θ. Alors*

$$\mathrm{EQM}_\theta(\hat{\theta}) = \mathrm{var}_\theta(\hat{\theta}) \geq \mathcal{I}(\theta)^{-1}.$$

Preuve. Comme $\hat{\theta}$ est sans biais, on a

$$\theta = \mathbb{E}_\theta(\hat{\theta}) = \int \hat{\theta}(x) V(\theta, x) d\nu(x) \ , \ \forall \theta \in \Theta.$$

La régularité du modèle entraîne en dérivant par rapport à θ,

$$1 = \int \hat{\theta}(x) \nabla V(\theta, x) d\nu(x) = \int \big(\hat{\theta}(x) - \theta\big) \nabla V(\theta, x) d\nu(x) = \int \big(\hat{\theta}(x) - \theta\big) S_\theta(x) V(\theta, x) d\nu(x)$$

avec $S_\theta(.) = \nabla V(\theta, .)/V(\theta, .)$ et où on a utilisé que $\int \nabla V(\theta, x) d\nu(x) = 0$. Par l'inégalité de Cauchy-Schwarz,

$$1 \leq \int \big(\hat{\theta}(x) - \theta\big)^2 V(\theta, x) d\nu(x) \ . \int S_\theta(x)^2 V(\theta, x) d\nu(x) = \mathrm{var}_\theta\left(\hat{\theta}\right) \mathcal{I}(\theta),$$

d'où le résultat. □

Le résultat reste valable en dimension supérieure: si $\Theta \subseteq \mathbb{R}^p$, alors tout estimateur sans biais $\hat{\theta}$ de θ vérifie

$$\mathrm{var}_\theta\left(\hat{\theta}\right) \geq \mathcal{I}(\theta)^{-1}$$

où l'inégalité entre matrices $A \geq B$ signifie que $A - B$ est semi définie positive. Ici, on a donc pour tout $a \in \mathbb{R}^p$,

$$\mathrm{var}_\theta\left(a^\top \theta\right) = a^\top \mathrm{var}_\theta\left(\hat{\theta}\right) a \geq a^\top \mathcal{I}(\theta)^{-1} a.$$

Pour le montrer, il faut adapter la preuve à une combinaison linéaire quelconque $a^\top \theta$ avec $a \in \mathbb{R}^p$. Ce résultat se généralise également à tout paramètre $\psi = \psi(\theta)$.

Corollaire 3.19. *Si $\psi : \Theta \to \mathbb{R}^q$ est une fonction différentiable, alors tout estimateur sans biais $\hat{\psi}$ de $\psi = \psi(\theta)$ vérifie*

$$\mathrm{var}_\theta\left(\hat{\psi}\right) \geq \nabla\psi(\theta)^\top \mathcal{I}(\theta)^{-1} \nabla\psi(\theta).$$

La preuve est identique. Le membre de droite est appelé borne de Cramer-Rao du paramètre $\psi(\theta)$.

Définition 3.20. *Dans un modèle régulier, un estimateur sans biais $\hat{\psi}$ d'un paramètre ψ est dit efficace si sa variance est égale à la borne de Cramer-Rao.*

Exercice. Soit un modèle exponentiel régulier canonique avec $V(\theta, X) = \exp(T(X)\theta - \zeta(\theta))$ la vraisemblance par rapport à ν et $\theta \in \mathbb{R}$.

1. Justifier que $\zeta(\theta) = \ln\left(\int \exp(T(x)\theta) d\nu(x)\right)$ et donc que ζ est deux fois dérivable.
2. Calculer $\mathbb{E}_\theta(T(X))$ et $\mathrm{var}_\theta(T(X))$ et fonction de ζ (et de ses dérivées).
3. Calculer l'information de Fisher du modèle puis la borne de Cramer-Rao du paramètre $\zeta'(\theta)$. Conclure.

Il n'existe qu'une famille restreinte de paramètres que l'on peut estimer efficacement.

Proposition 3.21. *Dans un modèle régulier avec $\Theta \subseteq \mathbb{R}^p$, si $\hat{\psi} = \hat{\psi}(X)$ est un estimateur efficace d'un paramètre réel $\psi = \psi(\theta) \in \mathbb{R}$, alors il existe une fonction g telle que*

$$\hat{\psi}(X) - \psi(\theta) \overset{ps}{=} g(\theta) \nabla\psi(\theta)^\top \mathcal{I}(\theta)^{-1} S_\theta(X).$$

Preuve. Reprenons le raisonnement de la preuve de l'inégalité de Cramer-Rao,

$$\nabla\psi(\theta) = \int \left(\hat{\psi}(x) - \psi(\theta)\right) S_\theta(x) V(\theta, x) d\nu(x).$$

Par l'inégalité de Cauchy-Schwarz, on a pour tout $a \in \mathbb{R}^p$,

$$\left(a^\top \nabla\psi(\theta)\right)^2 \leq \int \left(\hat{\psi}(x) - \psi(\theta)\right)^2 V(\theta, x) d\nu(x) \ . \int \left(a^\top S_\theta(x)\right)^2 V(\theta, x) d\nu(x),$$

avec égalité si et seulement si $\hat{\psi}(.) - \psi(\theta)$ est proportionnel à $a^\top S_\theta$. Pour obtenir l'inégalité de Cramer-Rao, il faut appliquer la majoration précédente à

$$a = \frac{\mathcal{I}(\theta)^{-1} \nabla\psi(\theta)}{\sqrt{\nabla\psi(\theta)^\top \mathcal{I}(\theta)^{-1} \nabla\psi(\theta)}}.$$

On déduit que $\hat{\psi}(X)$ est efficace ssi $\hat{\psi}(.) - \psi(\theta)$ est proportionnel à $\nabla\psi(\theta)^\top \mathcal{I}(\theta)^{-1} S_\theta(.)$ μ_θ-presque partout. $\square$

En dimension 1, $\Theta \subseteq \mathbb{R}$, ce résultat nous dit que si un estimateur sans biais $\hat{\psi}(X)$ est efficace, alors c'est une transformation affine du score $S_\theta(X)$. A transformation affine près, il n'existe donc au plus qu'un seul paramètre pour lequel on peut construire un estimateur efficace.

Quand il n'existe pas d'estimateur efficace, on peut parfois montrer l'existence d'un estimateur sans biais de variance finie uniformément minimale (minimale pour tout θ) parmi les estimateurs sans biais. On dira que c'est un estimateur sans biais *optimal*.

Proposition 3.22. *Un estimateur $\hat{\psi}$ de ψ est sans biais optimal si et seulement si il est de variance finie et pour toute statistique $T = T(X)$ telle que $\mathbb{E}_\theta(T) = 0$ et $\mathrm{var}_\theta(T) < \infty$ pour tout θ,*

$$\mathrm{cov}_\theta(\hat{\psi}, T) = \mathbb{E}_\theta(\hat{\psi}T) = 0.$$

Preuve. Soit $\hat{\psi}$ un estimateur sans biais optimal et T une statistique centrée de variance finie pour tout θ, alors $\hat{\psi} + \lambda T$ est un estimateur sans biais de ψ pour tout $\lambda \in \mathbb{R}$ et on a

$$\mathrm{var}_\theta(\hat{\psi} + \lambda T) = \mathrm{var}_\theta(\hat{\psi}) + \lambda^2 \mathrm{var}_\theta(T) + 2\lambda \,\mathrm{cov}_\theta(\hat{\psi}, T) \geq \mathrm{var}_\theta(\hat{\psi}).$$

Autrement dit $\lambda^2 \mathrm{var}_\theta(T) + 2\lambda \,\mathrm{cov}_\theta(\hat{\psi}, T) \geq 0$ pour tout $\lambda \in \mathbb{R}$ ce qui implique $\mathrm{cov}_\theta(\hat{\psi}, T) = 0$. Réciproquement, soit $\hat{\psi}$ un estimateur sans biais de ψ de variance finie. Pour tout estimateur sans biais $\tilde{\psi}$ de variance finie, $T = \hat{\psi} - \tilde{\psi}$ vérifie $\mathbb{E}_\theta(T) = 0$ et $\mathrm{var}_\theta(T) < \infty$. On a

$$\mathrm{var}_\theta(\tilde{\psi}) = \mathrm{var}_\theta(\hat{\psi} - T) = \mathrm{var}_\theta(\hat{\psi}) - 2\,\mathrm{cov}_\theta(\hat{\psi}, T) + \mathrm{var}_\theta(T).$$

On conclut que $\mathrm{var}_\theta(\hat{\psi}) \leq \mathrm{var}_\theta(\tilde{\psi})$ dès que $\mathrm{cov}_\theta(\hat{\psi}, T) = 0$. □

Corollaire 3.23. *Un estimateur sans biais optimal est unique presque sûrement pour tout θ.*

Preuve. Soient $\hat{\psi}, \tilde{\psi}$ deux estimateurs sans biais de ψ de variance uniformément minimale. On a

$$0 = \mathrm{var}_\theta(\tilde{\psi}) - \mathrm{var}_\theta(\hat{\psi}) = -2\,\mathrm{cov}_\theta(\hat{\psi}, \hat{\psi} - \tilde{\psi}) + \mathrm{var}_\theta(\hat{\psi} - \tilde{\psi}).$$

Or $\mathrm{cov}_\theta(\hat{\psi}, \hat{\psi} - \tilde{\psi}) = 0$ par la proposition précédente, d'où $\mathrm{var}_\theta(\hat{\psi} - \tilde{\psi}) = 0$ pour tout θ. □

Ces résultats ne sont informatifs que sous réserve d'existence d'au moins un estimateur sans biais de variance uniformément minimale. Cette condition n'est vérifiée que pour une famille assez restreinte de paramètre $\psi = \psi(\theta)$. En fait, même l'existence d'estimateur sans biais est loin d'être garantie.

Exercice. Montrer qu'il n'existe pas d'estimateur sans biais de $\psi(p) = 1/p$ dans le modèle d'échantillonage iid de Bernoulli $\mathcal{B}(p), p \in]0, 1[$.

Théorème 3.24 (Rao-Blackwell)**.** *Soit $\hat{\psi}$ un estimateur sans biais de variance finie d'un paramètre réel ψ et $S = S(X)$ une statistique exhaustive. Alors, l'estimateur $\hat{\psi}_S = \mathbb{E}_\theta\big(\hat{\psi}|S\big)$ est un estimateur sans biais de variance inférieure ou égale à celle de $\hat{\psi}$.*

L'exhaustivité de S est nécessaire pour que $\hat{\psi}_S$ soit bien un estimateur et donc ne dépende pas de θ.

Preuve. On vérifie facilement que $\hat{\psi}_S$ est bien un estimateur sans biais du fait que

$$\mathbb{E}_\theta\left(\hat{\psi}_S\right) = \mathbb{E}_\theta\left(\mathbb{E}_\theta\left(\hat{\psi}|S\right)\right) = \mathbb{E}_\theta\left(\hat{\psi}\right) = \psi.$$

De plus,

$$\mathrm{var}_\theta\left(\hat{\psi}\right) = \mathbb{E}_\theta\left(\hat{\psi} - \hat{\psi}_S + \hat{\psi}_S - \psi\right)^2 = \mathbb{E}_\theta\left(\hat{\psi} - \hat{\psi}_S\right)^2 + \mathbb{E}_\theta\left(\hat{\psi}_S - \psi\right)^2 + 2\mathbb{E}_\theta\left((\hat{\psi} - \hat{\psi}_S)(\hat{\psi}_S - \psi)\right).$$

Comme $\hat{\psi}_S - \psi$ est une fonction de S et $\hat{\psi}_S = \mathbb{E}_\theta(\hat{\psi}|S)$, on déduit $\mathbb{E}_\theta\left(\hat{\psi}(\hat{\psi}_S - \psi)\right) = \mathbb{E}_\theta\left(\hat{\psi}_S(\hat{\psi}_S - \psi)\right)$ et donc

$$\mathrm{var}_\theta\left(\hat{\psi}\right) = \mathrm{var}_\theta\left(\hat{\psi}_S\right) + \mathbb{E}_\theta\left(\hat{\psi} - \hat{\psi}_S\right)^2 \geq \mathrm{var}_\theta\left(\hat{\psi}_S\right). \qquad \square$$

Le théorème de Rao-Blackwell permet d'améliorer un estimateur sans biais si l'on connaît une statistique exhaustive mais ne garantit pas que celui-ci soit optimal. Par exemple, on sait que $S(X) = X$ est trivialement une statistique exhaustive mais que pour tout estimateur sans biais $\hat{\psi}$, la "Rao-Blackwellisation" $\mathbb{E}(\hat{\psi}|S) = \hat{\psi}$ ne modifie pas l'estimateur. En revanche, l'optimalité est garantie si la statistique exhaustive S est complète.

Théorème 3.25 (Lehmann-Scheffé)**.** *Soit $\hat{\psi}$ un estimateur sans biais de $\psi \in \mathbb{R}$ de variance finie et $S = S(X)$ une statistique exhaustive complète, alors $\hat{\psi}_S = \mathbb{E}_\theta(\hat{\psi}|S)$ est l'unique (presque sûrement) estimateur sans biais optimal.*

Preuve. Soit $\tilde{\psi}$ un autre estimateur sans biais de ψ de variance finie et $\tilde{\psi}_S = \mathbb{E}(\tilde{\psi}|S)$. On sait que $\hat{\psi}_S$ et $\tilde{\psi}_S$ sont des estimateurs sans biais de ψ et donc que $\mathbb{E}_\theta\left(\hat{\psi}_S - \tilde{\psi}_S\right) = 0$. Comme S est complète, cela implique que $\hat{\psi}_S = \tilde{\psi}_S$ presque sûrement pour tout θ, en particulier $\mathrm{var}_\theta(\hat{\psi}_S) = \mathrm{var}_\theta(\tilde{\psi}_S)$. Par le thèorème de Rao-Blackwell, $\tilde{\psi}$ ne peut donc pas avoir une variance inférieure à celle de $\hat{\psi}_S$. $\square$

Si l'on connaît une statistique exhaustive complète S dans le modèle, un estimateur sans biais est donc optimal si et seulement si il peut s'écrire comme une fonction de S presque sûrement.

Exercice. Soient $X = (X_1, ..., X_n)$ iid de loi de Bernoulli $\mathcal{B}(p), p \in]0, 1[$.

1. Sachant que $S(X) = \sum_{i=1}^n X_i$ est exhaustive complète, montrer que $\hat{p} = \frac{1}{n}\sum_{i=1}^n X_i$ est l'estimateur sans biais optimal de p.
2. Montrer que X_1 est un estimateur sans biais de p. En déduire $\mathbb{E}(X_1|S)$.

4 Estimation ponctuelle

4.1 Méthode des moments

Les moments d'une variable aléatoire réelle X sont donnés par

$$m_p := \mathbb{E}(X^p)\ ,\ p = 1, 2, ...$$

Le moment d'ordre p n'est défini que si la quantité $\mathbb{E}\left(|X|^p\right)$ est finie. On peut bien sûr généraliser cette définition à des moments p non-entiers.

Proposition 4.1. *Si X admet un moment d'ordre $p > 0$, alors X admet également un moment d'ordre q pour tout $q \in [0, p[$.*

Preuve. Soit $q \in [0, p[$, la fonction $\phi : x \mapsto x^{p/q}$ est convexe sur $[0, +\infty[$ puisque $p/q > 1$. Par l'inégalité de Jensen,

$$\mathbb{E}\big(|X|^p\big) = \mathbb{E}\Big(\big(|X|^q\big)^{p/q}\Big) \geq \Big(\mathbb{E}\big(|X|^q\big)\Big)^{p/q}$$

d'où le résultat. □

Les moments sont faciles à estimer à partir d'un échantillon iid $X_1, ..., X_n$. En effet, par la loi forte des grands nombres, on sait que si m_p existe alors

$$\widehat{m}_p := \frac{1}{n} \sum_{i=1}^{n} X_i^p$$

converge presque sûrement vers m_p quand $n \to \infty$. Si de plus, X_1 admet un moment d'ordre $2p$, alors le TCL est également valable et on a

$$\sqrt{n}\big(\widehat{m}_p - m_p\big) \xrightarrow[n\to\infty]{loi} \mathcal{N}\big(0, m_{2p} - m_p^2\big).$$

Remarque 4.2. *L'estimateur $\widehat{m}_p$ correspond au moment d'ordre p d'une variable aléatoire de fonction de répartition F_n (la fonction de répartition empirique). On l'appelle le moment empirique d'ordre p.*

Supposons maintenant que la loi des observations appartienne à un modèle paramétrique $\mathcal{M} = \{\mu_\theta \ , \ \theta \in \Theta\}$ pour lequel chacun des paramètres $\theta = (\theta_1, ..., \theta_k)$ s'exprime en fonction des moments m_p. Les estimateurs des moments $\widehat{m}_p$ permettent donc d'estimer naturellement les paramètres du modèle par "injection", c'est-à-dire en remplaçant les moments théoriques m_p par leurs versions empiriques $\widehat{m}_p$ dans l'expression de $\theta = \theta(m_1, m_2, ...)$.

Exemples.

1. Soit $X_1, ..., X_n$ iid dans le modèle Gaussien $\mathcal{N}(\theta, \sigma^2)$. Comme θ est le moment d'ordre 1: $m_1 = \mathbb{E}(X_i)$, un estimateur des moments de θ est

$$\widehat{\theta} = \widehat{m}_1 = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

   De plus, on remarque que $\sigma^2 = m_2 - m_1^2$. On déduit un estimateur de σ^2 par la méthode des moments:

$$\widehat{\sigma}^2 = \widehat{m}_2 - \widehat{m}_1^2 = \frac{1}{n} \sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n} \sum_{i=1}^{n} X_i\right)^2.$$

   par la loi forte des grands nombres (et continuité de la fonction $(x, y) \mapsto x - y^2$ pour le second).

2. Soit un échantillon iid $X_1, ..., X_n$ de loi de Poisson de paramètre $\lambda > 0$ inconnu. On sait que l'espérance de la loi de Poisson est égale au paramètre, on peut donc construire un estimateur de λ par

$$\hat{\lambda} = \widehat{m}_1 = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Poussons le vice plus loin en remarquant que λ est également la variance dans le modèle de Poisson: $\lambda = m_2 - m_1^2$. On déduit un second estimateur de λ:

$$\widetilde{\lambda} = \widehat{m}_2 - \widehat{m}_1^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2.$$

Ces deux estimateurs sont consistants.

3. On rappelle que l'espérance et la variance d'une variable aléatoire X de loi gamma $\Gamma(a,b)$ sont données par

$$\mathbb{E}(X) = \frac{a}{b} \quad \text{et} \quad \mathrm{var}(X) = \frac{a}{b^2}.$$

Soit $X_1, ..., X_n$ iid de loi gamma $\Gamma(a,b)$ avec $a, b > 0$. Pour estimer les paramètres a, b inconnus du modèle, on les exprime en fonction des moments en résolvant un système à (ici) deux équations

$$\begin{cases} m_1 = \dfrac{a}{b} \\ m_2 - m_1^2 = \dfrac{a}{b^2} \end{cases} \iff \begin{cases} a = \dfrac{m_1^2}{m_2 - m_1^2} \\ b = \dfrac{m_1}{m_2 - m_1^2} \end{cases}$$

On déduit des estimateurs $\widehat{a}, \widehat{b}$ par la méthode des moments en remplaçant les moments théoriques m_1, m_2 par les versions empiriques (on remarquera que $\widehat{m}_2 - \widehat{m}_1^2$ est la variance empirique et donc strictement positive presque sûrement à partir de $n \geq 2$).

Comme on a pu le constater dans les exemples, il n'y a pas qu'une façon de construire des estimateurs par la méthode des moments, puisque l'expression d'un paramètre θ en fonction de moments n'est en général pas unique. En pratique, on choisira généralement l'expression issue des premiers moments par simplicité, mais aussi parce que les moments empiriques d'ordre petit ont une variance plus faible. Pour estimer k paramètres qui ne sont pas liés, il faut au minimum k moments et résoudre un système d'équations.

Proposition 4.3. *On suppose l'existence du moment d'ordre k. Soit un paramètre $\theta = g(m_1, ..., m_k)$ avec $g : \mathbb{R}^k \to \mathbb{R}$ une fonction continue. L'estimateur par la méthode des moments*

$$\widehat{\theta} := g\left(\widehat{m}_1, ..., \widehat{m}_k\right)$$

converge presque sûrement vers $\theta = g(m_1, ..., m_k)$.

Preuve. Par la LFGN, $\hat{m}_j$ converge presque sûrement vers m_j pour tout $j = 1, ..., k$. Comme une intersection finie d'événements de probabilité 1 est de probabilité 1, le vecteur $(m_1, ..., m_k)$ converge lui-aussi presque sûrement. Par continuité de g (la continuité au point $(m_1, ..., m_k)$ suffit), $\widehat{\theta} \xrightarrow{ps} \theta$. □

Remarque 4.4. *Alors que les moments empiriques sont bien des estimateurs sans biais des moments théoriques, rien n'indique que l'estimateur des moments $\widehat{\theta} = g\left(\hat{m}_1, ..., \hat{m}_k\right)$ reste sans biais dès que la fonction g n'est pas linéaire. C'est le cas par exemple de la variance empirique*

$$\widehat{\sigma}^2 = \widehat{m}_2 - \widehat{m}_1^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2,$$

qui n'est pas sans biais (mais l'est asymptotiquement).

Sous des conditions de régularité supplémentaires sur g, on peut étudier plus en détail le comportement asymptotique d'un estimateur de la forme $\widehat{\theta} := g(\widehat{m}_1, ..., \widehat{m}_k)$. On utilise pour cela la Δ-méthode qui est basé sur un développement limité de g au point $(m_1, ..., m_k)$.

Proposition 4.5 (Δ-méthode)**.** *Soit $(X_n)_{n\in\mathbb{N}}$ une suite de vecteurs aléatoires de $\mathbb{R}^k$ qui converge en probabilité vers le vecteur constant $m = (m_1, ..., m_k)^\top$ et $(a_n)_{n\in\mathbb{N}}$ une suite positive qui tend vers $+\infty$ tels que*

$$a_n(X_n - m) \xrightarrow[n\to\infty]{loi} Y.$$

Si $g(.) = (g_1(.), ..., g_l(.)) : \mathbb{R}^k \to \mathbb{R}^l$ est une fonction de classe $\mathcal{C}^1$ dans un voisinage $\mathcal{V}$ de m, alors

$$a_n(g(X_n) - g(m)) \xrightarrow[n\to\infty]{loi} \nabla g(m)^\top Y,$$

où $\nabla g(m) \in \mathbb{R}^{l\times k}$ est la matrice des dérivées partielles $\big(\partial_{m_j} g_i(m)\big)_{i=1,...,l, j=1,...,k}$.

Preuve. On applique la formule de Taylor-Lagrange à l'ordre 1, si $X_n(\omega) \in \mathcal{V}$:

$$g(X_n(\omega)) = g(m) + \nabla g(m)^\top (X_n(\omega) - m) + o\big(\|X_n(\omega) - m\|\big).$$

On obtient

$$a_n\big(g(X_n(\omega)) - g(m)\big) = \nabla g(m)^\top a_n(X_n(\omega) - m) + o\big(a_n\|X_n(\omega) - m\|\big).$$

Or $\lim_{n\to\infty} \mathbb{P}(X_n \in \mathcal{V}) = 1$ par la convergence en probabilité de X_n vers m, ce qui permet de conclure. □

Dans le cas qui nous intéresse, la Δ-méthode permet de montrer la normalité asymptotique de tout estimateur de la forme $\widehat{\theta} := g(\widehat{m}_1, ..., \widehat{m}_k)$ pour g une fonction régulière.

Corollaire 4.6. *On suppose que X_i admet un moment d'ordre $2k$. Soit $\theta = g(m_1, ..., m_k)$ où g est de classe $\mathcal{C}^1$ dans un voisinage de $m = (m_1, ..., m_k)^\top$, l'estimateur des moments $\widehat{\theta} := g(\widehat{m}_1, ..., \widehat{m}_k)$ vérifie*

$$\sqrt{n}\big(\widehat{\theta} - \theta\big) \xrightarrow[n\to\infty]{loi} \mathcal{N}(0, \nabla g(m)^\top \Sigma \nabla g(m)),$$

où $\Sigma_{ij} = m_{i+j} - m_i m_j$ pour $i, j = 1, ..., k$.

4.2 Maximum de vraisemblance

On appelle estimateur du maximum de vraisemblance toute variable aléatoire qui maximise $\theta \mapsto V(\theta, X)$. En général, rien ne garantit l'unicité ni même l'existence du maximum de vraisemblance.

Exemples

1. Modèle binomial: on observe le résultat de n lancés de dés $X = (X_1, ..., X_n) \in \{0, 1\}^n$. Les X_i sont des réalisations indépendantes d'une variable de Bernoulli de paramètre p_0 inconnu. La probabilité d'observer l'échantillon est de

$$\prod_{i=1}^n p_0^{X_i}(1 - p_0)^{1-X_i} = p_0^S(1 - p_0)^{n-S},$$

où $S = \sum_{i=1}^n X_i$. La vraie valeur p_0 étant inconnue, on peut écrire cette probabilité comme une fonction de p:

$$V(p, X) = p^S (1-p)^{n-S}$$

qui est la probabilité d'observer l'échantillon $X = (X_1, ..., X_n)$ si les observations suivent la loi de Bernoulli de paramètre p (qui n'est autre que la vraisemblance du modèle). L'estimateur du maximum de vraisemblance est la valeur $\hat{p}$ qui maximise la probabilité $V(p, X)$,

$$\hat{p} = \arg\max_{p \in [0,1]} V(p, X).$$

On note $v(p, X) = \log V(p, X)$ la log-vraisemblance, maximiser $V(., X)$ revient à maximiser $v(., X)$. La fonction $p \mapsto v(p, X)$ est strictement concave sur $[0, 1]$, le maximiseur est donc l'unique solution de la condition du premier ordre

$$\frac{\partial}{\partial \theta} v(\hat{p}, X) = \frac{S}{\hat{p}} - \frac{n-S}{1-\hat{p}} = 0 \Longleftrightarrow \hat{p} = \frac{S}{n}.$$

Dans ce modèle l'estimateur du maximum de vraisemblance est donc défini comme la variable aléatoire $\hat{p} = \frac{1}{n} \sum_{i=1}^n X_i$.

2. Modèle de Poisson: la vraisemblance pour un modèle de Poisson $\mathcal{M} = \{\mathcal{P}(\lambda), \lambda > 0\}$ est donnée par

$$V(\lambda, X) = e^{-n\lambda} \frac{\lambda^S}{\prod_{i=1}^n X_i!}.$$

Maximiser $V(\lambda, X)$ revient à maximiser $\lambda \mapsto \log(\lambda) S - n\lambda$. Le maximiseur est

$$\hat{\lambda} = \frac{1}{n} \sum_{i=1}^n X_i.$$

3. Loi uniforme: on considère le modèle $\mathcal{M} = \{\mathcal{U}(0, \theta), \theta > 0\}$, qui sont des lois de densités

$$f_\theta(x) = \begin{cases} 1/\theta & \text{si } x \in [0, \theta] \\ 0 & \text{sinon.} \end{cases}$$

La vraisemblance du modèle est donnée par

$$V(\theta, X) = \begin{cases} \theta^{-n} & \text{si } \max\{X_1, ..., X_n\} \leq \theta \\ 0 & \text{sinon.} \end{cases}$$

La fonction $\theta \mapsto \theta^{-n}$ est décroissante sur $[0, +\infty[$, on déduit donc que la vraisemblance est maximale pour $\hat{\theta} = \max\{X_1, ..., X_n\}$.

4. Loi normale: ...

Le maximum de vraisemblance est stable par transformation des observations et du paramètre.

Proposition 4.7. *Soit un modèle paramétrique $\mathcal{M}$ de vraisemblance $V(\theta, X), \theta \in \Theta$ dominé par ν la mesure de Lebesgue ou de comptage et $\hat{\theta}$ un estimateur du maximum de vraisemblance de θ.*

i) Soit $g : \mathcal{X} \to \mathcal{Y}$ un $\mathcal{C}^1$ diffeomorphisme, $\hat{\theta}$ est un estimateur du maximum de vraisemblance pour les observations $Y = g(X)$.

ii) Soit $\psi : \Theta \to \Psi$ une fonction injective, $\hat{\psi} = \psi(\hat{\theta})$ est un estimateur du maximum de vraisemblance de $\psi = \psi(\theta)$.

Preuve. Si ν est la mesure de Lebesgue, la densité de Y sous μ_θ est donnée par

$$V_Y(\theta, y) = \frac{1}{J_g \circ g^{-1}(y)} V(\theta, g^{-1}(y)) \ , \ y \in \mathcal{Y},$$

où $J_g(.)$ est le Jacobien de g. Evaluée en $y = Y$, on trouve

$$V_Y(\theta, Y) = \frac{1}{J_g \circ g^{-1}(Y)} V(\theta, g^{-1}(Y)) = \frac{1}{J_g(X)} V(\theta, X).$$

De même, si ν est la mesure de comptage, alors $V_Y(\theta, Y) = V(\theta, X)$. Dans les deux cas, maximiser $V_Y(., Y)$ revient à maximiser $V(., X)$. Pour *ii)*, il suffit de remarquer que $V(\theta, X) = V(\psi^{-1}(\psi(\theta)), X)$ où $V(\psi^{-1}(.), X)$, définie sur Ψ, est la vraisemblance du modèle paramétré par ψ. Comme ψ est injective, on a bien

$$V(\psi^{-1}(\hat{\psi}), X) = V(\hat{\theta}, X) = \sup_{\theta \in \Theta} V(\theta, X) = \sup_{\psi \in \Psi} V(\psi^{-1}(\psi), X). \qquad \square$$

Dans le cas *ii)*, si ψ n'est pas injective, le résultat reste vrai en définissant la vraisemblance du modèle paramétré par ψ par

$$V_\psi(\psi, X) = \sup_{\theta : \psi(\theta) = \eta} V(\theta, X).$$

Dans le cas iid, sous des conditions (raisonnables mais techniques) de régularité, on peut montrer que l'estimateur du maximum de vraisemblance existe et est unique et consistant. On note f_θ la densité de X_1 sous θ par rapport à ν et $\ell_\theta = \ln(f_\theta)$, la log-vraisemblance s'écrit donc

$$v(\theta, X) = \sum_{i=1}^{n} \ell_\theta(X_i) \ , \ \theta \in \Theta.$$

On suppose $\ell_{\theta'}$ intégrable sous μ_θ pour tout $\theta, \theta' \in \Theta$ et on note

$$\forall \theta, \theta' \in \Theta \ , \ \xi_\theta(\theta') := \mathbb{E}_\theta(\ell_{\theta'}(X_1)) = \int \ell_{\theta'}(x) f_\theta(x) d\nu(x).$$

Sous μ_θ, la log-vraisemblance renormalisée $\hat{\xi} := v(., X)/n$ est un estimateur sans biais de ξ_θ qui converge simplement presque sûrement par la loi forte des grands nombres

$$\forall \theta' \in \Theta \ , \ \hat{\xi}(\theta') = \frac{1}{n} \sum_{i=1}^{n} \ell_{\theta'}(X_i) \xrightarrow[n \to \infty]{\mu_\theta - \text{p.s.}} \xi_\theta(\theta').$$

Lemme 4.8. *La fonction ξ_θ atteint son maximum en θ.*

Preuve. Pour tout $\theta' \in \Theta$, on a par l'ingélité de Jensen appliquée au logarithme (concave):

$$\xi_\theta(\theta') = \int \ln(f_\theta) f_\theta d\nu + \int \ln\left(\frac{f_{\theta'}}{f_\theta}\right) f_\theta d\nu \leq \int \ln(f_\theta) f_\theta d\nu + \ln\left(\int \frac{f_{\theta'}}{f_\theta} f_\theta d\nu\right) = \xi_\theta(\theta). \qquad \square$$

Par passage à la limite, on s'attend donc à ce qu'un maximiseur de $\hat{\xi}$ (et donc de la vraisemblance) converge vers θ si celui-ci est l'unique maximiseur de ξ. Une condition nécessaire, que l'on supposera

par la suite, est donc que le modèle soit fortement identifiable au sens où pour toute suite $(\theta_k)_{k\in\mathbb{N}}$ de Θ,

$$\lim_{k\to\infty} \xi_\theta(\theta_k) = \xi_\theta(\theta) \implies \lim_{k\to\infty} \theta_k = \theta,$$

ce qui peut s'écrire de manière équivalente

$$f_{\theta_k} \xrightarrow[k\to\infty]{\nu-\text{p.p.}} f_\theta \implies \lim_{k\to\infty} \theta_k = \theta.$$

Cette condition dépend essentiellement de la paramétrisation du modèle et est vérifiable en pratique si on impose qu'elle soit vraie pour tout θ.

Théorème 4.9. *Soit $\hat{\theta}$ un estimateur du maximum de vraisemblance de θ. Si la famille de fonctions $\{\ell_{\theta'}, \theta' \in \Theta\}$ vérifie la loi des grands nombres uniforme sous μ_θ (c'est une classe de Glivenko-Cantelli sous μ_θ):*

$$\forall \theta \in \Theta \ , \ \|\hat{\xi} - \xi_\theta\|_\infty = \sup_{\theta' \in \Theta} \left|\hat{\xi}(\theta') - \xi_\theta(\theta')\right| \xrightarrow[n\to\infty]{\mu_\theta - p.s.} 0,$$

alors $\hat{\theta}$ converge presque sûrement vers θ.

Preuve. Soit $\hat{\theta}$ un estimateur du maximum de vraisemblance, on a par l'inégalité triangulaire

$$|\xi_\theta(\hat{\theta}) - \xi_\theta(\theta)| \leq |\xi_\theta(\hat{\theta}) - \hat{\xi}(\hat{\theta})| + |\hat{\xi}(\hat{\theta}) - \xi_\theta(\theta)| \leq 2\|\hat{\xi} - \xi_\theta\|_\infty \underset{n\to\infty}{\longrightarrow} 0.$$

Par la condition d'identifiabilité forte, $\hat{\theta}$ converge donc vers θ, μ_θ-p.s. □

On montre maintenant la normalité asymptotique de l'estimateur du maximum de vraisemblance. On supposera en plus que le modèle est régulier et on note $s_\theta = \nabla \ell_\theta$ et $h_\theta = \nabla^\top \nabla \ell_\theta$ les fonctions gradient et la matrice Hessienne de ℓ_θ par rapport à θ.

Théorème 4.10. *Sous les conditions du théorème précédent, si $\mathcal{I}(\theta) = -\int h_\theta f_\theta d\nu$ (l'information de Fisher pour une observation) est inversible, alors*

$$\sqrt{n}\big(\hat{\theta} - \theta\big) \xrightarrow[n\to\infty]{loi} \mathcal{N}\big(0, \mathcal{I}(\theta)^{-1}\big).$$

Preuve. Comme le modèle est régulier, un estimateur du maximum de vraisemblance vérifie les conditions du premier ordre $\nabla\hat{\xi}(\hat{\theta}) = 0$. Par un développement de Taylor:

$$0 = \nabla\hat{\xi}(\hat{\theta}) = \frac{1}{n}\sum_{i=1}^n s_{\hat{\theta}}(X_i) = \frac{1}{n}\sum_{i=1}^n s_\theta(X_i) + \left(\frac{1}{n}\sum_{i=1}^n h_\theta(X_i)\right)(\hat{\theta} - \theta) + o\|\hat{\theta} - \theta\|.$$

Or $\frac{1}{n}\sum_{i=1}^n h_\theta(X_i)$ converge presque sûrement vers $-\mathcal{I}(\theta)$ et est donc inversible avec probabilité 1 asymptiquement, auquel cas

$$\sqrt{n}(\hat{\theta} - \theta) = -\left(\frac{1}{n}\sum_{i=1}^n h_\theta(X_i)\right)^{-1} \frac{\sqrt{n}}{n}\sum_{i=1}^n s_\theta(X_i) + o\big(\sqrt{n}\|\hat{\theta} - \theta\|\big).$$

On remarque de plus que $\frac{\sqrt{n}}{n}\sum_{i=1}^n s_\theta(X_i) \xrightarrow{loi} \mathcal{N}\big(0, \mathcal{I}(\theta)\big)$ par le théorème central limit. On conclut par le lemme de Slutsky. □

Dans un modèle régulier, l'estimateur du maximum de vraisemblance de θ est donc asymptotiquement Gaussien de variance l'inverse de l'information de Fisher. On dira alors qu'il est asymptotiquement efficace puisque sa loi limite quand $n \to \infty$ est d'espérance θ (il est asymptotiquement sans biais) et de variance la borne de Cramer-Rao. Ce résultat reste valable pour un estimateur du maximum de vraisemblance d'un paramètre $\psi = \psi(\theta)$ avec ψ différentiable. En effet, dans ce cas l'estimateur du maximum de vraisemblance $\hat{\psi} = \psi(\hat{\theta})$ vérifie par la Δ-méthode

$$\sqrt{n}\big(\hat{\psi} - \psi\big) \xrightarrow[n\to\infty]{loi} \mathcal{N}\big(0, \nabla\psi(\theta)^\top \mathcal{I}(\theta)^{-1} \nabla\psi(\theta)\big),$$

où on reconnaît ici aussi comme variance la borne de Cramer-Rao du paramètre ψ. Attention, la convergence en loi n'implique pas la convergence des moments.

5 Estimation par région de confiance

Jusqu'à présent, nous avons vu des méthodes d'estimation qui ont pour objectif, étant donné un paramètre inconnu $\psi = \psi(\theta)$, de proposer une valeur en fonction des observations: un estimateur. Plutôt que de se limiter à une valeur, un point de vue différent consiste à proposer un ensemble de valeurs, associé à une probabilité d'appartenance fixée à l'avance.

On se placera dans un modèle paramétrique $\mathcal{M} = \{\mu_\theta : \theta \in \Theta\}$ avec X le vecteur aléatoires des observations à valeurs dans un espace $\mathcal{X}$ (typiquement $\mathcal{X} = \mathbb{R}^n$ dans les cas d'observations iid réelles). On considère un paramètre $\psi = \psi(\theta)$ et on note $\Psi = \psi(\Theta)$ l'image de Θ par $\psi(.)$.

5.1 Définitions générales

Définition 5.1. *On appelle région de confiance d'un paramètre ψ un sous-ensemble aléatoire $R = R(X) \subseteq \Psi$ fonction des observations. Le niveau de confiance $1-\alpha \in [0,1]$ de R est donné par*

$$1 - \alpha := \inf_{\theta\in\Theta} \mathbb{P}_\theta(R \ni \psi).$$

Si $R(X)$ est un intervalle pour toute valeur possible de X, alors on parle d'intervalle de confiance.

L'idée d'une région de confiance est de fournir, étant donnée une marge d'erreur α fixée à l'avance, un ensemble de valeurs possibles pour le paramètre ψ qui contiennent la vraie valeur avec probabilité supérieure ou égale à $1-\alpha$ (et donc avec une probabilité d'erreur inférieure ou égale à α). Comme cette probabilité dépend de la vraie valeur de θ qui est inconnue, α est définie comme la probabilité d'erreur dans le pire cas possible: $\alpha = \sup_{\theta\in\Theta} \mathbb{P}_\theta(R \not\ni \psi)$. Dans certaines situations idéales, la probabilité d'appartenance à la région de confiance ne dépend pas de θ auquel cas le niveau de confiance $1-\alpha$ est atteint pour tout θ. On dira que le niveau de confiance est *exactement égal* à $1-\alpha$ dans ce cas.

Définition 5.2. *Une fonction $g : \mathcal{X} \times \Psi \to \mathbb{R}$ est pivotale si la loi de $g(X,\psi)$ ne dépend pas de θ.*

Toute fonction pivotale permet de construire une région de confiance.

Proposition 5.3. *Si g est pivotale, alors pour tout ensemble $G_\alpha \subset \mathbb{R}$ tel que $\mathbb{P}_\theta\big(g(X,\psi) \in G_\alpha\big) = 1-\alpha$, la région*

$$R_\alpha(X) := \{\psi \in \Psi : g(X,\psi) \in G_\alpha\}$$

est une région de confiance de niveau exactement égal à $1-\alpha$.

Preuve. Il suffit de remarquer que $\mathbb{P}_\theta\big(\psi \in R_\alpha(X)\big) = \mathbb{P}_\theta\big(g(X,\psi) \in G_\alpha\big) = 1-\alpha$. $\square$

On peut construire un ensemble G_α vérifiant $\mathbb{P}_\theta\big(g(X,\psi) \in G_\alpha\big) = 1-\alpha$ précisément parce que g est pivotale et donc que la loi de $g(X,\psi)$ est connue (ne dépend pas de θ).

Lorsqu'on ne sait pas construire de fonction pivotale, on peut se contenter d'une fonction asymptotiquement pivotale $g_n(X,\psi)$ qui converge en loi sous μ_θ pour tout $\theta \in \Theta$ et dont la loi limite ne dépend pas de θ. Cela permet de construire une région de confiance asymptotique.

Exemples.

1. Modèle d'échantillonage iid exponentiel de paramètre $\lambda > 0$: $g(\lambda, X) = \sum_{i=1}^n X_i$ est pivotale de loi $\Gamma(1,n)$.

2. On cherche un intervalle de confiance sur le paramètre p d'un modèle de Bernoulli. On sait que $\sqrt{n}(\hat{p}-p)$ est asymptotiquement Gaussien de moyenne 0 et variance $p(1-p)$. On a donc par le lemme de Slutsky,
$$\sqrt{n}\frac{\hat{p}-p}{\sqrt{\hat{p}(1-\hat{p})}} \xrightarrow[n\to\infty]{loi} \mathcal{N}(0,1).$$
On obtient l'intervalle de confiance asymptotique de niveau α:
$$\Big[\hat{p} - q_{1-\frac{\alpha}{2}}\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}\ ,\ \hat{p} + q_{1-\frac{\alpha}{2}}\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}\Big].$$
On obtient une région de confiance différente (mais toujours de niveau α asymptotiquement) en considérant la statistique $\sqrt{n}(\hat{p}-p)/\sqrt{p(1-p)}$ (exercice).

3. Modèle de Poisson de paramètre λ: $\sqrt{n}(\hat{\lambda}-\lambda)/\sqrt{\lambda}$ est asymptotiquement pivotale de loi limite $\mathcal{N}(0,1)$, pour $\hat{\lambda} = \frac{1}{n}\sum_{i=1}^n X_i$.

5.2 Le modèle Gaussien

On suppose ici que les observations sont iid de loi normale $\mathcal{N}(m,\sigma^2)$. On veut construire un intervalle de confiance sur la moyenne m et/ou la variance σ^2.

1. *Moyenne avec variance connue.* On suppose que les observations $X_1, ..., X_n$ sont distribuées suivant une loi normale de variance σ^2 connue et de moyenne m inconnue. Dans ce modèle on sait que
$$\hat{m} = \frac{1}{n}\sum_{i=1}^n X_i \sim \mathcal{N}\Big(m, \frac{\sigma^2}{n}\Big).$$
La loi de l'estimateur est connue en fonction de m. En particulier, on peut se ramener à une variable aléatoire normale centrée réduite par
$$g(X,m) = \sqrt{n}\ \frac{\hat{m}-m}{\sigma} \sim \mathcal{N}(0,1).$$
La fonction $g(X,m) = \sqrt{n}(\hat{m}-m)/\sigma$ est donc pivotale puisque sa loi ne dépend pas de m. On insiste sur le fait que l'on n'observe pas une réalisation de $g(X,m)$ puisque c'est une fonction du paramètre inconnu m. La symétrie de la densité Gaussienne nous permet d'écrire
$$\mathbb{P}(-q_{1-\frac{\alpha}{2}} \le g(m,X) \le q_{1-\frac{\alpha}{2}}) = 1-\alpha,$$

où $q_{1-\frac{\alpha}{2}}$ est le quantile d'ordre $1-\frac{\alpha}{2}$ de la loi normale centrée réduite. On cherche maintenant à isoler m dans l'expression de $g(X,m)$ de manière à en obtenir un encadrement. On écrit

$$\mathbb{P}\Big(-q_{1-\frac{\alpha}{2}} \leq \sqrt{n}\,\frac{\hat{m}-m}{\sigma} \leq q_{1-\frac{\alpha}{2}}\Big) = \mathbb{P}\Big(\hat{m} - \frac{q_{1-\frac{\alpha}{2}}\,\sigma}{\sqrt{n}} \leq m \leq \hat{m} + \frac{q_{1-\frac{\alpha}{2}}\,\sigma}{\sqrt{n}}\Big) = 1-\alpha.$$

On remarque que les bornes de l'intervalles sont fonctions des observations et donc connues. On obtient ainsi un intervalle de confiance pour m avec une marge d'erreur de α qui correspond à la probabilité que la vraie valeur du paramètre ne soit pas dans l'intervalle.

2. *Moyenne avec variance inconnue.* Ici, on ne peut pas utiliser la statistique précédente car elle dépend de σ^2. On remplace donc la variance par l'estimateur sans biais

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\hat{m})^2.$$

On utilise alors les propriétés suivantes

- $(n-1)\hat{\sigma}^2/\sigma^2$ suit une loi du Khi-2 à $n-1$ degrés de liberté, notée $\chi^2(n-1)$
- $\hat{\sigma}^2$ et $\hat{m}$ sont indépendantes (conséquence du théorème de Basu).

On déduit que la fonction

$$g(X,m) = \sqrt{n}\,\frac{\hat{m}-m}{\hat{\sigma}} = \frac{\sqrt{n}(\hat{m}-m)/\sigma}{\hat{\sigma}/\sigma} \overset{loi}{=} \frac{\mathcal{N}(0,1)}{\sqrt{\chi^2(n-1)/(n-1)}} \Big\} \text{ indépendantes}$$

est pivotale. Sa loi est appelée loi de Student à $n-1$ degrés de liberté et notée $\mathcal{T}(n-1)$. On peut alors construire un intervalle de confiance de la même façon que pour le cas σ^2 connu remplaçant les quantiles de la loi normale standard par ceux de la loi de Student à $n-1$ degrés de liberté.

3. *Variance avec moyenne connue.* On sait $(X_i-m)/\sigma$ est de loi $\mathcal{N}(0,1)$. Si m est connu, la fonction

$$g(X,\sigma^2) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-m)^2$$

est donc pivotale, de loi $\chi^2(n)$. Soit $k_{\alpha,n}$ le quantile d'ordre α de la loi du Khi deux à n degrés de liberté, qui vérifie $\mathbb{P}(\chi^2(n) \geq k_{\alpha,n}) = 1-\alpha$. On a

$$\mathbb{P}(k_{\frac{\alpha}{2},n} \leq g(X,\sigma^2) \leq k_{1-\frac{\alpha}{2},n}) = \mathbb{P}\left(\frac{\sum_{i=1}^{n}(X_i-m)^2}{k_{1-\frac{\alpha}{2},n}} \leq \sigma^2 \leq \frac{\sum_{i=1}^{n}(X_i-m)^2}{k_{\frac{\alpha}{2},n}}\right) = 1-\alpha.$$

4. *Variance avec moyenne inconnue.* On a vu que $\hat{\sigma}^2$ vérifie que $(n-1)\hat{\sigma}^2/\sigma^2 \sim \chi^2(n-1)$. On obtient donc un intervalle de confiance en utilisant les quantiles de la loi $\chi^2(n-1)$ (et non $\chi^2(n)$ comme dans l'exemple précédent).

6 Tests paramétriques

En statistique, choisir un modèle revient à fixer des conditions sur la loi des observations. Lorsque l'on souhaite vérifier si ces conditions sont raisonnables, on peut utiliser un *test statistique*. Formellement, un test est une variable aléatoire construite à partir des observations qui prend la valeur 1 si on rejette l'hypothèse, ou 0 si on ne rejette pas l'hypothèse. La décision de rejeter ou non l'hypothèse est toujours sensible à l'aléa des observations, ce qui nous oblige à fixer à l'avance une marge d'erreur α qui correspond à la probabilité de rejeter l'hypothèse alors qu'elle est vraie. Pour tester une hypothèse H_0, on construit une variable aléatoire dont la loi est connue si H_0 est vraie. Si la valeur observée de cette variable est peu vraisemblable par rapport à sa loi sous H_0, on décidera de rejeter l'hypothèse.

6.1 Définitions générales

On se place dans le cadre d'un modèle paramétrique $\mathcal{M} = \{\mu_\theta : \theta \in \Theta\}$. Faire un test sur la loi dont sont issues les observations X revient à tester un ensemble de valeurs du paramètre θ (on ne remet pas en doute que la vraie loi appartient au modèle). L'hypothèse que l'on veut tester, appelée *hypothèse nulle*, s'écrit donc $H_0 : \theta \in \Theta_0$ où Θ_0 est un sous-ensemble connu de Θ. L'hypothèse complémentaire $H_1 : \theta \in \Theta_1 := \Theta \setminus \Theta_0$ est appelée *hypothèse alternative*. Il peut arriver de considérer des alternatives différentes avec $\Theta_1 \neq \Theta \setminus \Theta_0$ mais on aura toujours $\Theta_0 \cap \Theta_1 = \emptyset$.

Définition 6.1. *Un test d'une hypothèse $H_0 : \theta \in \Theta_0$ contre une alternative $H_1 : \theta \in \Theta_1$ est une statistique $\Phi = \Phi(X)$ de loi de Bernoulli qui vaut 1 si on rejette H_0 et 0 sinon. Le niveau α du test est la probabilité (maximale) de rejeter H_0 alors qu'elle est vraie*

$$\alpha = \sup_{\theta \in \Theta_0} \mathbb{P}_\theta(\Phi = 1).$$

Si la probabilité de rejeter H_0 est contante pour tout $\theta \in \Theta$, on dit que le test est de niveau exactement égal à α. Le niveau α représente la marge d'erreur autorisée si on décide de rejeter H_0, auquel cas on dira qu'on rejette H_0 au risque α. Il est en général fixé à l'avance et le test Φ est construit de manière à être de niveau α.

Définition 6.2. *Soit Φ un test, la région critique R (ou région de rejet) est l'ensemble des réalisations possibles qui conduisent à rejeter H_0: $R = \{x \in \mathcal{X} : \Phi(x) = 1\}$. On a donc*

$$\Phi = \mathbb{1}\{X \in R\}.$$

Similairement, la région d'acceptation A est le complémentaire: $A = \mathcal{X} \setminus R$.

Certains niveaux α peuvent être impossibles à atteindre exactement. Par exemple si X ne peut prendre qu'un nombre fini de valeurs, l'ensemble $\mathcal{P}(\mathcal{X})$ des parties de $\mathcal{X}$ est fini et il est donc impossible de trouver un ensemble R tel que $\mathbb{P}_\theta(X \in R) = \alpha$ pour tout $\alpha \in (0, 1)$. On peut alors choisir de se contenter du niveau atteignable le plus proche ou construire un test dit randomisé. Pour un test randomisé paramétrique, l'espace $\mathcal{X}$ est partitionné en trois sous-ensembles R, B, A avec:

- R la région critique qui vérifie $\mathbb{P}_\theta(X \in R) < \alpha$ pour tout $\theta \in \Theta_0$.
- B une zone intermédiaire pour laquelle si $X \in B$, on rejette H_0 avec une certaine probabilité γ calculée de manière à ce que $\sup_{\theta \in \Theta}\{\mathbb{P}_\theta(X \in R) + \gamma \mathbb{P}_\theta(X \in B)\} = \alpha$. Conditionnellement au fait que $X \in B$, la décision de rejeter H_0 se fait indépendamment de X.

- A la région d'acceptation $A = \mathcal{X} \setminus (R \cup B)$.

Cette approche permet d'obtenir un niveau α artificiellement (en ajoutant de l'alée dès que $X \in B$) même lorsque ce niveau impossible à atteindre pour un test fonction de X uniquement.

En pratique, on dispose souvent d'une collection croissante de régions critiques $R_\alpha, \alpha \in]0,1[$ pour chaque niveau, qui est croissante au sens où

$$\forall \alpha' > \alpha : R_{\alpha'} \supset R_\alpha.$$

Autrement dit, rejeter l'hypothèse H_0 au niveau α implique de la rejeter également pour tous les niveaux supérieurs. On dispose donc d'un continuum de tests $\Phi_\alpha, \alpha \in]0,1[$ qui donnent des événements $\{\Phi_\alpha = 1\}, \alpha \in]0,1[$ croissants. C'est typiquement le cas quant la région critique peut s'exprimer à partir d'une statistique de test $T = T(X) \in \mathbb{R}$ sous la forme

$$R_\alpha = \{x \in \mathcal{X} : T(x) > s_\alpha\}$$

où s_α est un seuil adapté.

Définition 6.3. *On appelle* p-value *du test le plus petit niveau α pour lequel on rejette H_0 au vu de la réalisation de Φ_α:*

$$\text{p-value} = \inf\{\alpha > 0 : \Phi_\alpha(X) = 1\}.$$

On rejettera donc l'hypothèse H_0 si la p-value est supérieure au niveau du test. L'écart entre la p-value et α quantifie la marge avec laquelle on accepte H_0 dans le cas où elle est plus grande que α et la marge avec laquelle on rejette H_0 dans le cas contraire.

Proposition 6.4. *Pour un test exact, la p-value est une variable aléatoire de loi uniforme sur $[0,1]$ sous H_0.*

Preuve. On suppose que H_0 est vraie. Soit $u \in]0,1[$,

$$\mathbb{P}(\text{p-value} \leq u) = \mathbb{P}\Big(\inf\{\alpha > 0 : \Phi_\alpha(X) = 1\} \leq u\Big) = \mathbb{P}\Big(\bigcup_{0<\alpha\leq u} \{\Phi_\alpha(X) = 1\}\Big).$$

Comme les événements $\{\Phi_\alpha(X) = 1\}, \alpha \in [0,1]$ sont croissants et de probabilité α, on déduit que $\mathbb{P}(\text{p-value} \leq u) = \lim_{\alpha \to u} \alpha = u$. □

Exemple. Dans un modèle Gaussien iid $\mathcal{N}(\theta, 1), \theta \in \mathbb{R}$, on veut tester si les observations sont centrées, à savoir $H_0 : \theta = 0$. L'hypothèse alternative est $H_1 : \theta \neq 0$. On peut estimer le paramètre θ par $\hat{\theta} = \frac{1}{n}\sum_{i=1}^n X_i$. Si H_0 est vraie, on sait que $T := \sqrt{n}\hat{\theta}$ suit une loi normale standard.
On peut raisonner en disant que l'hypothèse H_0 n'est crédible que si $\sqrt{n}\hat{\theta}$ est "aussi proche de zéro que ce qu'on peut attendre d'une réalisation de loi normale standard", on décide donc de rejeter H_0 si la valeur observée de T est supérieure en valeur absolue à un certain seuil k (il existe bien sûr d'autres tests possibles). Ce seuil est déterminé par le niveau α qui correspond à la probabilité d'erreur tolérée (probabilité de rejeter H_0 alors qu'elle est vraie, typiquement $\alpha = 0.05$). Etant donné α, on choisit donc le seuil de manière à ce que la probabilité de rejeter H_0 alors qu'elle est vraie soit égale à α:

$$\mathbb{P}_0(|T| \geq k) = \mathbb{P}(|\mathcal{N}(0,1)| \geq k) = \alpha \iff k = q_{1-\frac{\alpha}{2}},$$

où $q_.$ désigne le quantile de la loi normale standard. La conclusion du test est donc de rejeter H_0 au risque α de se tromper si $|T| > q_{1-\frac{\alpha}{2}}$ et de ne pas rejeter sinon. La p-value du test correspond à la probabilité obtenue en remplaçant le seuil s par la valeur obtenue $|T|$ (grossièrement, la probabilité d'obtenir ce qu'on a observé sous H_0)

$$\text{p-value} = \mathbb{P}(|\mathcal{N}(0,1)| \geq |T|) = 2(1 - F(|T|))$$

où F est la fonction de répartition de la $\mathcal{N}(0,1)$. On peut vérifier que si H_0 est vraie et donc que $T \sim \mathcal{N}(0,1)$, la p-value est bien de loi uniforme sur $[0,1]$.

Définition 6.5. *Soit Φ_α un test de niveau α de l'hypothèse $H_0 : \theta \in \Theta_0$ contre $H_1 : \theta \in \Theta_1$. La puissance du test, notée $1 - \beta(.)$, est la fonction qui à $\theta \in \Theta_1$ associe la probabilité de rejeter H_0:*

$$1 - \beta(\theta) = \mathbb{P}_\theta\big(\Phi_\alpha = 1\big) \ , \ \theta \in \Theta_1.$$

La quantité $\beta(\theta)$ est la probabilité de ne pas rejeter H_0 si X est issu de la loi μ_θ. On dira qu'un test est consistant si la puissance $1 - \beta(\theta)$ tend vers 1 quand $n \to \infty$, pour tout $\theta \in \Theta_1$.

Il y a deux types d'erreurs possibles lors d'un test statistique:

- Rejeter H_0 alors qu'elle est vraie constitue l'erreur de première espèce, de probabilité inférieure ou égale au niveau α par construction. Elle est fixée à l'avance par le statisticien.
- Ne pas rejeter H_0 sous l'alternative $\theta \in \Theta_1$ constitue l'erreur de deuxième espèce, de probabilité $\beta(\theta)$. Un test sera d'autant meilleur (plus puissant) que l'erreur de deuxième espèce est faible, à un niveau α fixé, pour les différentes valeur de $\theta \in \Theta_1$. La puissance d'un test dépend donc fortement de l'hypothèse alternative H_1.

Exercice. Soit Φ un test de niveau α et fonction puissance $1 - \beta(.)$.

1. Soit $S = S(X)$ une statistique, montrer que la loi conditionnelle de Φ sachant S est une loi de Bernoulli de paramètre $\mathbb{E}_\theta(\Phi|S)$.
2. Réciproquement, si S est exhaustive, montrer que toute variable aléatoire de Bernoulli dont la loi conditionnelle sachant S est de paramètre $\mathbb{E}_\theta(\Phi|S)$ est un test de niveau α et puissance $1 - \beta(.)$ (préciser en quoi l'exhaustivité est nécessaire).
3. Commenter.

Définition 6.6. *Un test de niveau α de l'hypothèse $H_0 : \theta \in \Theta_0$ contre $H_1 : \theta \in \Theta_1$ est dit uniformément plus puissant (UPP) parmi les tests de niveau α si sa puissance $1-\beta(\theta)$ est supérieure ou égale à celle de n'importe quel autre test de niveau α, pour tout $\theta \in \Theta_1$.*

Il n'existe pas toujours de test UPP car la relation d'ordre sur les fonctions puissance n'est pas totale (sauf si Θ_1 est un singleton).

Exemple. Reprenons l'exemple Gaussien iid sur la moyenne θ pour tester $H_0 : \theta = 0$ contre $H_1 : \theta \neq 0$ avec $T = \sum_{i=1}^n X_i/\sqrt{n}$. Le test bilatéral $\Phi_\alpha = \mathbb{1}\{|T| \geq q_{1-\frac{\alpha}{2}}\}$ a pour puissance

$$1 - \beta(\theta) = \mathbb{P}_\theta(|T| \geq q_{1-\frac{\alpha}{2}}) = \mathbb{P}_\theta(T \geq q_{1-\frac{\alpha}{2}}) + \mathbb{P}_\theta(T \leq q_{\frac{\alpha}{2}}) \ , \theta \neq 0.$$

ou encore, pour F est la fonction de répartition de la loi normale standard,

$$1 - \beta(\theta) = F(q_{1-\frac{\alpha}{2}} - \theta) + F(q_{\frac{\alpha}{2}} - \theta) \ , \ \theta \neq 0.$$

On peut montrer que ce test n'est pas UPP. Par exemple le test unilatéral $\Phi_\alpha^+ = \mathbb{1}\{T \geq q_{1-\alpha}\}$ est bien de niveau α et a pour puissance

$$1 - \beta^+(\theta) = F(q_{1-\alpha} - \theta) \ , \ \theta \neq 0,$$

On voit que $1 - \beta^+(\theta) > 1 - \beta(\theta)$ pour tout $\theta > 0$, le test Φ_α^+ est donc plus puissant pour ces valeurs. En revanche, Φ_α^+ est très peu puissant pour $\theta < 0$ où $1 - \beta^+(\theta) < \alpha$ (on rejette moins facilement H_1 pour ces valeurs de θ que sous H_0). De même, le test unilatéral $\Phi_\alpha^- = \mathbb{1}\{T \leq q_\alpha\}$, lui aussi de niveau α, a des propriétés symétriques, il est plus puissant pour des valeurs négatives de θ et très peu puissant pour $\theta > 0$. Aucun de ces trois tests n'est donc UPP.
La situation est différente si on considère $H_0 : \theta = 0$ contre $H_1 : \theta > 0$, ce qui revient à se placer dans le modèle $\{\mathcal{N}(\theta, 1) : \theta \geq 0\}$. Ici, le test unilatéral Φ_α^+ est le plus puissant des trois pour tout $\theta > 0$. Ce test est même UPP, ce qui est plus difficile à montrer.

Définition 6.7. *Un test de niveau α de l'hypothèse $H_0 : \theta \in \Theta_0$ contre $H_1 : \theta \in \Theta_1$ est dit sans biais si $1 - \beta(\theta) \geq \alpha$ pour tout $\theta \in \Theta_1$.*

Autrement dit, un test est sans biais si on a plus de chances de rejeter l'hypothèse alternative H_1 lorsqu'elle est fausse que lorsqu'elle est vraie. C'est évidemment une propriété souhaitable mais qui peut être difficile à obtenir. Dans l'exemple Gaussien précédent, pour tester $H_0 : \theta = 0$ contre $H_1 : \theta \neq 0$, seul le test Φ_α est sans biais parmi les trois tests considérés.

Exercice. Montrer qu'un test UPP est sans biais.

Définition 6.8. *Un test de niveau α de l'hypothèse $H_0 : \theta \in \Theta_0$ contre $H_1 : \theta \in \Theta_1$ est dit uniformément plus puissant sans biais (UPPSB) s'il est sans biais et uniformément plus puissant parmi les tests sans biais de niveau α.*

6.2 Test du rapport de vraisemblance

Pour tester l'hypothèse $H_0 : \theta \in \Theta_0$ contre $H_1 : \theta \in \Theta_1$, on peut naturellement comparer les vraisemblances et rejeter H_0 s'il existe une valeur de θ dans Θ_1 pour laquelle la vraisemblance est "significativement" plus grande que sur Θ_0.

Définition 6.9. *Pour $H_0 : \theta \in \Theta_0$ contre $H_1 : \theta \in \Theta_1$, la statistique*

$$\tau = \tau(X) = \frac{\sup_{\theta \in \Theta_0} V(\theta, X)}{\sup_{\theta \in \Theta_1} V(\theta, X)},$$

est appelée statistique du rapport de vraisemblance. On rejette H_0 si τ est inférieur à un seuil k_α calculé en fonction du niveau α du test.

On peut définir de manière équivalente la statistique

$$\tau' = \frac{\sup_{\theta \in \Theta_0} V(\theta, X)}{\sup_{\theta \in \Theta} V(\theta, X)}.$$

pour le test du rapport de vraisemblance.

Exercice. Montrer que si des tests $\Phi_\alpha = \mathbb{1}\{\tau < k_\alpha\}$ et $\Phi'_\alpha = \mathbb{1}\{\tau' < s'_\alpha\}$ basés respectivement sur les statistiques du rapport de vraisemblances τ et τ' sont de niveau α, alors les seuils k_α, s'_α et les régions de rejet correspondantes sont égales.

Proposition 6.10. *Pour toute statistique exhaustive S, le test du rapport de vraisemblance peut s'écrire comme une fonction de S.*

Preuve. C'est une conséquence immédiate du critère de factorisation de Fisher-Neyman. □

Exercice. Dans le modèle iid Gaussien $\mathcal{N}(\theta, 1), \theta \in \mathbb{R}$, exprimer la région de rejet du test du rapport de vraisemblance de l'hypothèse $H_0 : \theta = 0$ contre $H_1 : \theta \neq 0$ en fonction de la statistique exhaustive $S(X) = \frac{1}{n}\sum_{i=1}^n X_i$.

On s'intéresse maintenant à la situation où on veut tester une valeur particulière du paramètre. Le résultat fondamental suivant donne le test le plus puissant si l'hyptothèse alternative est également réduite à un singleton.

Théorème 6.11 (Lemme de Neyman-Pearson)**.** *Pour tester $H_0 : \theta = \theta_0$ contre $H_1 : \theta = \theta_1$, le test du rapport de vraisemblance*

$$\Phi := \mathbb{1}\{\tau < k_\alpha\} = \mathbb{1}\big\{V(\theta_0, X) < k_\alpha V(\theta_1, X)\big\}$$

est le test le plus puissant parmi les tests de niveau inférieur ou égal à $\alpha = \mathbb{P}_{\theta_0}(\tau < k_\alpha)$.

Preuve. Soit $R = \{x \in \mathcal{X} : \tau(x) < k_\alpha\}$ la région de rejet du test. La puissance du test, ici réduite à une valeur car Θ_1 est un singleton, est donnée par $1 - \beta = \mathbb{P}_{\theta_1}(X \in R) = \mathbb{P}_{\theta_1}(\tau < k_\alpha)$. Soit R' la région de rejet d'un test différent Φ' de niveau $\mathbb{P}_{\theta_0}(X \in R') \leq \alpha$ et de puissance $1 - \beta' = \mathbb{P}_{\theta_1}(X \in R')$. On a

$$1 - \beta = \mathbb{P}_{\theta_1}(X \in R) = \mathbb{P}_{\theta_1}(X \in R \cap R') + \mathbb{P}_{\theta_1}(X \in R \setminus R'),$$

avec

$$\mathbb{P}_{\theta_1}(X \in R \setminus R') = \int_{R \setminus R'} V(\theta_1, x) d\nu(x) \geq \int_{R \setminus R'} \frac{1}{k_\alpha} V(\theta_0, x) d\nu(x) = \frac{1}{k_\alpha}\mathbb{P}_{\theta_0}(X \in R \setminus R').$$

Or $\mathbb{P}_{\theta_0}(X \in R \setminus R') = \mathbb{P}_{\theta_0}(X \in R) - \mathbb{P}_{\theta_0}(X \in R') + \mathbb{P}_{\theta_0}(X \in R' \setminus R)$. Comme Φ' est de niveau inférieur ou égal à α, on déduit que $\mathbb{P}_{\theta_0}(X \in R \setminus R') \geq \mathbb{P}_{\theta_0}(X \in R' \setminus R)$. De plus,

$$\mathbb{P}_{\theta_0}(X \in R' \setminus R) = \int_{R' \setminus R} V(\theta_0, x) d\nu(x) \geq \int_{R' \setminus R} k_\alpha V(\theta_1, x) d\nu(x) = k_\alpha \mathbb{P}_{\theta_1}(X \in R' \setminus R).$$

En combinant les inégalités précédentes, on trouve

$$\mathbb{P}_{\theta_1}(X \in R \setminus R') \geq \mathbb{P}_{\theta_1}(X \in R' \setminus R)$$

et donc que $1 - \beta \geq 1 - \beta'$. □

Corollaire 6.12. *Soit $\Theta_1 \subset \Theta$ et $\theta_0 \in \Theta \setminus \Theta_1$, s'il existe une statistique réelle $S = S(X)$ telle que $V(\theta_0, X)/V(\theta, X) = h_\theta(S)$ où $h_\theta(.)$ est une fonction strictement croissante pour tout $\theta \in \Theta_1$, alors le test du rapport de vraisemblance est UPP parmi les tests de niveau inférieur ou égal pour tester $H_0 : \theta = \theta_0$ contre $H_1 : \theta \in \Theta_1$.*

Preuve. Ici,

$$\tau(X) = \frac{V(\theta_0, X)}{\sup_{\theta\in\Theta_1} V(\theta, X)} = \inf_{\theta\in\Theta_1} h_\theta(S).$$

On a donc comme région de rejet

$$R = \Big\{ \inf_{\theta\in\Theta_1} h_\theta(S) < k_\alpha \Big\} = \bigcap_{\theta\in\Theta_1} \{h_\theta(S) < k_\alpha\} = \bigcap_{\theta\in\Theta_1} \{S < h_\theta^{-1}(k_\alpha)\} = \Big\{ S < \inf_{\theta\in\Theta_1} h_\theta^{-1}(k_\alpha) \Big\}.$$

Le seuil $s_\alpha = \inf_{\theta\in\Theta_1} h_\theta^{-1}(k_\alpha)$ ne dépend pas de $\theta \in \Theta_1$, le test est donc identique au test de $H_0 : \theta = \theta_0$ contre $H_1 : \theta = \theta_1$ pour tout $\theta_1 \in \Theta_1$. Or, ce test est le plus puissant par le lemme de Neyman-Pearson. □

Si θ est un paramètre réel, les conditions du corollaire sont vérifiées pour $\Theta_1 = \{\theta : \theta > \theta_0\}$ si le rapport $V(\theta, X)/V(\theta', X)$ est une fonction strictement croissante de S pour tout $\theta < \theta'$. Dans ce cas, le test du rapport de vraisemblance est UPP pour tester $H_0 : \theta \leq \theta_0$ contre $H_1 : \theta > \theta_0$. Si de plus le modèle est une famille exponentielle avec $V(\theta, X) = G(X)H(\theta)\exp\big(T(X)\eta(\theta)\big)$, alors

$$\frac{V(\theta, X)}{V(\theta', X)} = \frac{H(\theta)}{H(\theta')} \exp\big(T(X)(\eta(\theta) - \eta(\theta'))\big)$$

et le test du rapport de vraisemblance pour $H_0 : \theta \leq \theta_0$ contre $H_1 : \theta > \theta_0$ est donc UPP si $\eta(.)$ est strictement monotone (avec $S = T$ si $\eta(.)$ est croissante et $S = -T$ si elle est décroissante).

Exercice. Trouver une décomposition de la forme $V(\theta_0, X)/V(\theta, X) = h_\theta(S)$ avec $h_\theta(.)$ croissante pour tout $\theta > \theta_0$ dans les modèles suivants:

- Modèle Gaussien $\mathcal{N}(\theta, 1)$.
- Modèle de Cauchy centré en θ.
- Modèle binomial $\mathcal{B}(k, \theta)$ avec k fixé.
- Modèle de Poisson $\mathcal{P}(\theta)$.

Théorème 6.13 (Théorème de Wilks)**.** *Dans un modèle d'échantillonage paramétrique iid régulier avec $\Theta \subseteq \mathbb{R}^p$, sous les conditions qui garantissent la normalité et efficacité asymptotique de l'estimateur du maximum de vraisemblance $\hat{\theta}$, la statistique du rapport de vraisemblance τ pour $H_0 : \theta = \theta_0$ contre $H_1 : \theta \neq \theta_0$ vérifie sous H_0:*

$$-2\ln(\tau) \xrightarrow[n\to\infty]{loi} \chi^2(p).$$

Preuve. Par continuité de la vraisemblance τ et τ' sont égales ici et on a

$$-2\ln(\tau) = -2\ln\left(\frac{V(\theta_0, X)}{V(\hat{\theta}, X)}\right) = -2\big(v(\theta_0, X) - v(\hat{\theta}, X)\big) = -2\sum_{i=1}^n (\ell_{\theta_0}(X_i) - \ell_{\hat{\theta}}(X_i)),$$

où $\ell_\theta(.) = \ln f_\theta(.)$, le logarithme de la densité par rapport à la mesure dominante ν. Par un développement limité à l'ordre 2,

$$-2\ln(\tau) = 2\big(\hat{\theta} - \theta_0\big)^\top \sum_{i=1}^n s_{\theta_0}(X_i) + \big(\hat{\theta} - \theta_0\big)^\top \left(\sum_{i=1}^n h_{\theta_0}(X_i)\right)\big(\hat{\theta} - \theta_0\big) + o\big(\|\hat{\theta} - \theta_0\|^2\big).$$

Sous H_0, on sait que $\sqrt{n}\big(\hat{\theta}-\theta_0\big)$ converge en loi vers une loi $\mathcal{N}(0, \mathcal{I}(\theta_0)^{-1})$ tandis que $\frac{1}{n}\sum_{i=1}^n h_{\theta_0}(X_i)$ converge presque sûrement vers $\mathcal{I}(\theta_0)$ par la LFGN. On déduit que

$$-2\ln(\tau) = \|\sqrt{n}\mathcal{I}(\theta_0)^{-1/2}\big(\hat{\theta}-\theta_0\big)\|^2 + o_P(1),$$

où $\sqrt{n}\mathcal{I}(\theta_0)^{-1/2}\big(\hat{\theta}-\theta_0\big)$ converge en loi vers un vecteur Gaussien standard de $\mathbb{R}^p$. $\square$

Le théorème de Wilks permet de construire un test asymptotiquement de niveau α directement pour $H_0 : \theta = \theta_0$ à partir de la statistique du rapport de vraisemblance. Le seuil est alors un quantile de la loi du Khi-2 correspondante et ne dépend donc que de la dimension du modèle.

Le théorème de Wilks peut s'appliquer plus généralement à des hypothèses de la forme $H_0 : \theta \in \Theta_0$ contre $H_1 : \theta \notin \Theta_0$. Dans ce cas, si $p_0 < p$ désigne la dimension de Θ_0, on a sous les mêmes conditions de régularité

$$-2\ln(\tau) \xrightarrow[n\to\infty]{loi} \chi^2(p-p_0).$$

Le test d'indépendance sur des couples (X_i, Y_i) de variables aléatoires iid discrètes à support fini en est un cas particulier (exercice).