Statistique Mathématique

O. Wintenberger

Préambule

Ce polycopié s'adresse aux étudiants ayant suivi un cours d'intégration, un premier cours de probabilité et un premier cours de statistique. Les notions d'algèbre linéaire et de probabilités requises sont dans le fascicule "Rappels utiles au cours de statistique mathématique" disponible à l'adresse http ://wintenberger.fr/ens.
La première partie présente les notions fondamentales de l'inférence statistique, à savoir les notions d'échantillonnage, d'empirique et d'information. La seconde partie traite de l'estimation, ponctuelle ou par intervalle de confiance. La troisième partie introduit la notion de test statistique.

Table des matières

Introduction

La science des statistiques comporte 2 aspects :

1. **Les statistiques descriptives** qui consistent à synthétiser, résumer, structurer l'information contenue dans les données (cf monographie d'"Introduction à la méthode statistique" de Goldfarb et Pardoux),
2. **La statistique mathématique** qui consiste à traduire en langage mathématique la démarche d'inférence statistique.

L'inférence statistique :

L'inférence statistique est le fait de fournir à partir d'une propriété observée dans des cas particuliers des caractéristiques de la propriété en général. Par essence cette démarche est risquée et s'oppose à la démarche déductive (non risquée) qui applique les caractéristiques d'une propriété en général à des cas particuliers et qu'on rencontre généralement en mathématique.

Sous des hypothèses probabilistes spécifiques issues de la modélisation du problème, il est possible de traduire l'inférence statistique en langage mathématique. Dans ce cours on étudie le traitement mathématique de deux démarches inférentielles spécifiques : l'estimation et le test.

Première partie

Premiers principes de l'inférence statistique

Chapitre 1

L'échantillon aléatoire

A partir de l'observation d'une propriété sur des cas particuliers (en nombre fini) le statisticien infère des caractèristiques de la propriété en général. La statistique mathématique se divise selon deux approches : l'approche bayésienne qui suppose que cette propriété est aléatoire et l'approche fréquentiste qui suppose que cette propriété est déterministe. Nous nous restreignons dans ce cours au cadre fréquentiste.

Deux cas de figure sont possibles :
- Soit la propriété est observée sur un sous ensemble de taille n d'une population mère de taille finie N avec $N >> n$,
- Soit la propriété est observée sur un ensemble fini d'expériences issues du renouvellement de la même expérience.

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

1.1 Population de taille finie

Soit E un ensemble, que nous appellerons population mère (des individus, un parc automobile), contenant un nombre fini N d'éléments. Le statisticien s'intéresse plus particulièrement à une propriété X de la population (l'âge, le prix), appelée propriété statistique. L'objectif du statisticien est de déterminer les principales caractéristiques de X.

S'il est possible d'effectuer un recensement, c'est-à-dire interroger ou inspecter tous les éléments de E, les caractéristiques de la propriété X sont parfaitement connues. Si on note e_i chaque élément de E, $E = \{e_1, \ldots, e_N\}$, on observe alors $(x_1, \ldots, x_N)$ l'ensemble des valeurs de X mesurées sur les éléments de E.

L'inférence statistique n'est pas utile dans le cas d'un recensement mais lorsque X est observée uniquement sur un sous-ensemble de E (pour des raisons de coût, de commodité,..) notée $\mathcal{E}_n$ de taille $n << N$: $\mathcal{E}_n = \{e_{i_1}, \ldots, e_{i_n}\}$ où $1 \leq i_k \leq N$ et $1 \leq k \leq n$. Nous supposons avoir procédé à la sélection de l'échantillon $\mathcal{E}_n$ de manière aléatoire et avec remise : on sélectionne au hasard un élément de E puis il est "remis" dans la population et peut être de nouveau sélectionné ultérieurement. De fait, il peut y avoir un couple (k, k') tel que $i_k = i_{k'}$. On est alors dans le cas d'un tirage aléatoire avec remise. Il est clair qu'il existe dans ce cas $N^n/n!$ différentes possibilités pour choisir $\mathcal{E}_n$. L'inférence statistique est effectuée à partir d'observations de la propriété X sur $\mathcal{E}_n$. On note $X_1, \ldots, X_n$ les valeurs de X correspondant aux éléments de $\mathcal{E}_n$. Ce sont des valeurs aléatoires car $\mathcal{E}_n$ a été tiré aléatoirement et le vecteur $(X_1, \ldots, X_n)$ est l'échantillon.

1.2 Expérience renouvelable

Les modèles où la population est de taille finie ne couvrent pas toutes les situations. Prenons le cas de la propriété X égale à "la fréquence, mesurée en minutes, de départ du métro de la ligne 2 à la station Porte Dauphine". Il est clair que X est une variable aléatoire puisqu'on ne peut exactement prédire la fréquence. En revanche,on ne peut pas appliquer les notions de population finie et d'échantillonnage aléatoire ici car le nombre d'observations dépend du temps qu'on passe à observer le métro. On parle plutôt ici d'expérience que l'on peut renouveler théoriquement autant de fois que l'on veut.

On considère le cas d'une expérience aléatoire renouvelée plusieurs fois indépendamment. On note X la propriété statistique associé à l'expérience et dont les caractéristiques sont inconnues du statisticien. Alors X_1 correspond à la propriété X observée sur la première expérience. L'expérience est renouvelée n fois afin d'obtenir les observations $X_1, \ldots, X_n$ puis le statisticien infère à partir de ces données pour déduire des caractéristiques sur la propriété X.

1.3 L'échantillon

Afin de donner à l'échantillon un cadre mathématique commun, on suppose que la propriété X appartient à un espace euclidien $\mathcal{X}$ ($\mathbb{R}^q$ avec $q \geq 1$) muni de sa norme euclidienne $\|\cdot\|$. On suppose aussi que l'ensemble des caractéristiques de la propriété X sont décrites par une mesure de probabilité P inconnue. Alors X est un élément aléatoire (e.a.) à valeur dans $\mathcal{X}$ de loi P. C'est donc une application

mesurable de $(\Omega, \mathcal{A})$ dans $(\mathcal{X}, \mathcal{B})$, où $\mathcal{B}$ est la tribu des Boréliens et $(\Omega, \mathcal{A}, \mathbb{P})$ est l'ensemble des événements possibles muni d'une mesure de probabilité, vérifiant $\mathbb{P}(X \in B) = P(B)$ pour tout $B \in \mathcal{B}$.

Définition 1.3.1 *L'échantillon aléatoire* $(X_1, \ldots, X_n)$ *de taille* n *est le vecteur aléatoire à valeur dans l'espace produit* $(\mathcal{X}, \mathcal{B})^n = (\mathcal{X}^n, \mathcal{B}_n)$ *de loi* $P^{\otimes n}$ *où*

- $\mathcal{X}^n = \underbrace{\mathcal{X} \times \ldots \times \mathcal{X}}_{n \text{ fois}}$ *est le produit cartésien de l'espace* $\mathcal{X}$,
- $\mathcal{B}_n$ *est la tribu des Boréliens de* $\mathcal{X}^n$,
- $P^{\otimes n} = P \otimes \cdots \otimes P$ *le produit tensoriel de* P n*-fois.*

Pour tout $1 \leq i \leq n$ *la ième observation* X_i *est un e.a. de même loi* P *que* X*. Les observations sont indépendantes entre elles.*

On note $X_1, \ldots, X_n \sim P$ ou $\sim F$, F étant la fonction de répartition de X. Par définition du produit tensoriel, on a

$$P^{\otimes n}(B_1 \times \cdots \times B_n) = \prod_{j=1}^{n} P(B_j),$$

pour tout $B_1, \ldots, B_n \in \mathcal{B}$. Dans le cas continu où P admet une densité f (relativement à la mesure de Lebesgue), l'échantillon $(X_1, \ldots, X_n)$ admet aussi une densité donnée par la formule :

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

pour tout $x_1, \ldots, x_n \in \mathcal{X}$.

Remarque 1 *Dans le cas d'une population de taille finie* E*, étant donné que la propriété* X *prend les valeurs* $\{x_1, \ldots, x_N\}$ *de manière équitable, c'est à dire avec probabilité identique, on trouve*

$$P(X = x_l) = 1/N, \quad \forall\ l = 1, \ldots, N.$$

On appelle cette loi la loi Uniforme Discrète sur l'ensemble $\{x_1, \ldots, x_N\}$*. On note* $X_1, \ldots, X_n$ *les observations de* X *sur* $\mathcal{E}_n$*, un échantillon aléatoire de taille* n *de* E*. La notation* $X_1, \ldots, X_n$ *ne signifie en aucun cas que les* n *premiers éléments de la population ont été observés. on vérifie bien que* $X_1, \ldots, X_n \sim P$ *car le tirage avec remise assure que les observations sont iid.*

Définition 1.3.2 *Une réalisation* $(x_1, \ldots, x_n)$ *de l'échantillon aléatoire* $(X_1, \ldots, X_n)$ *est le résultat des mesures associées à un événement* $A \in \mathcal{A}$ *:*

$$(x_1, \ldots, x_n) = (X_1(A), \ldots, X_n(A)).$$

C'est un élément déterministe de $\mathcal{X}^n$*. La réalisation* x_i *de la* i*-ème observation sera appelée plus simplement la* i*-ème réalisation.*

Chapitre 2

La méthode empirique

Le statisticien veut inférer sur une caractéristique précise de la propriété statistique X à partir de l'échantillon $(X_1, \ldots, X_n)$. Cette caractéristique peut s'écrire comme une fonction φ de la loi inconnue P de X et s'écrit donc $\varphi(P)$. La méthode empirique consiste à substituer à P inconnue la mesure empirique P_n fournit à partir de l'échantillon $(X_1, \ldots, X_n)$ (donc connue) par la relation

$$P_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}$$

où δ_a est la mesure de Dirac au point $a \in \mathcal{X}$: $\mathbb{P}(\delta_a = a) = 1$.

Remarque 2 *La mesure empirique P_n est la loi uniforme discrète sur l'ensemble des observations $\{X_1, \ldots, X_n\}$.*

Ce chapitre étudie différentes caractéristiques empiriques $\varphi(P_n)$ correspondant à différents choix de φ, plus spécifiquement

- La moyenne empirique $\frac{1}{n} \sum_{i=1}^{n} X_i$ notée $\overline{X}_n$,
- La matrice de variance-covariance empirique $\frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X}_n)(X_i - \overline{X}_n)^T$ notée S_n^2.

Dans le cas réel $\mathcal{X} = \mathbb{R}$, on étudie aussi

- Le moment empirique d'ordre $r \in \mathbb{N}$, $\frac{1}{n} \sum_{i=1}^{n} X_i^r$ noté M_n^r,
- Le moment empirique centré d'ordre $r \in \mathbb{N}^*$, $\frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X}_n)^r$ noté $M_n^{r\prime}$,

– La fonction de répartition empirique notée F_n qui dans le cas $\mathcal{X} = \mathbb{R}$ est la fonction qui à $x \in \mathbb{R}$ associe la valeur $\frac{1}{n}\sum_{i=1}^{n} 1_{X_i \leq x}$.

Pour faire appel aux théorèmes limites probabilistes, on a besoin de la notion suivante :

Définition 2.0.3 *Dans le cas d'une expérience renouvelable, lorsqu'on suppose pouvoir renouveler l'expérience autant de fois que l'on veut, i.e. faire tendre $n \to \infty$, on parle du cadre asymptotique.*

Ce cadre idéal permet d'appliquer les théorèmes classiques de convergence tels que la LFGN et le TLC. Dans ce cours, on étudiera principalement les résultats de type TLC :

Définition 2.0.4 *Une suite de vecteurs aléatoires (X_i) vérifie un TLC lorsqu'il existe un vecteur gaussien centré N et un vecteur déterministe μ tel que*

$$\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{\mathcal{L}} N \qquad \textit{quand} \qquad n \to \infty.$$

Bien que la convergence en loi soit un mode de convergence plus faible que la convergence p.s., le TLC implique la LFGN

Proposition 2.0.1 *Si une suite de vecteurs aléatoires (X_n) satisfait le TLC alors $\overline{X}_n \xrightarrow{p.s.} \mu$ asymptotiquement.*

Démonstration : Sans perte de généralité on pose $\mu = 0$. On utilise le lemme de Borel-Cantelli qui assure que si la série $(\mathbb{P}(\|\overline{X}_n\| > \varepsilon))$ est convergente pour tout $\varepsilon > 0$ alors $\overline{X}_n \xrightarrow{p.s.} 0$ asymptotiquement. On raisonne par équivalence (cas $q = 1$)

$$\mathbb{P}(\|\overline{X}_n\| > \varepsilon) = \mathbb{P}(\sqrt{n}|\overline{X}_n| > \varepsilon\sqrt{n}) \sim_{n\to\infty} 2\int_{\varepsilon\sqrt{n}}^{\infty} \frac{1}{2\pi} \exp\Big(-\frac{x^2}{2\sigma^2}\Big) dx =: u_n.$$

Or, (u_n) est majorée pour n suffisamment grand par une suite convergente :

$$u_n \leq \sqrt{\frac{2}{\pi}} \int_{\varepsilon\sqrt{n}}^{\infty} \frac{1}{2\pi} \exp(-x) dx = \sqrt{\frac{2}{\pi}} \exp(-\varepsilon\sqrt{n}).$$

Par croissance comparée, (u_n) est une série convergente ainsi que $(\mathbb{P}(\|\overline{X}_n\| > \varepsilon))$ et le résultat souhaité découle du lemme de Borel-Cantelli.

2.1 La moyenne empirique

Définition 2.1.1 *La moyenne empirique de l'échantillon est l'e.a.*

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

Même pour cette caractéristique très simple (la moyenne), la loi de la version empirique $\overline{X}_n$ n'est pas connue pour tous choix possibles de P. Par contre, on peut calculer des caractéristiques de l'e.a. $\overline{X}_n$ telles que $\mathbb{E}(\overline{X}_n)$ et $\mathrm{Var}\,(\overline{X}_n)$ dans un cadre général.

2.1.1 L'espérance de la moyenne empirique

Proposition 2.1.1 *Si l'e.a. X est intégrable, i.e. la loi P est telle que $\mathbb{E}(\|X\|) = \int \|x\| dP(x) < \infty$ alors*

$$\mathbb{E}(\overline{X}_n) = \mu$$

où $\mu = \mathbb{E}(X) = \int x dP(x)$ est la moyenne.

Démonstration : Application immédiate de la linéarité de l'intégrale. □

Exemple 2.1.1 *Dans le cas d'une population E de taille finie N, on calcule*

$$\mu = \int x dP(x) = \sum_{\ell=1}^{N} x_\ell P(X = x_\ell) = \frac{1}{N}\sum_{j=1}^{N} x_j$$

et on obtient

$$\mathbb{E}(\overline{X}_n) = \frac{1}{N}\sum_{j=1}^{N} x_j = \overline{x}_N.$$

2.1.2 La matrice de variance-covariance de $\overline{X}_n$

Proposition 2.1.2 *Si l'e.a. X est de carré intégrable, i.e. la loi P est telle que $\mathbb{E}(\|X\|^2) = \int \|x\|^2 dP(x) < \infty$ alors*

$$Var\,(\overline{X}_n) = \frac{\Sigma^2}{n}$$

où $\Sigma^2 = \mathbb{E}((X - \mathbb{E}(X))(X - \mathbb{E}(X))^T) = \mathbb{E}(XX^T) - \mathbb{E}(X)\mathbb{E}(X)^T$ est la matrice de variance-covariance.

Démonstration : Les X_j étant des e.a. iid, on a

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

et $\mathrm{Var}\,(X_i) = \Sigma^2$ pour tout $\forall\ j = 1, \ldots, n$. □

Exemple 2.1.2 *Dans le cas d'une population E de taille finie N, on calcule*

$$\sigma^2 = \frac{1}{N}\sum_{j=1}^{N}(x_j - \mu)^2,$$

et on obtient

$$Var\,(\overline{X}_n) = \frac{\sum_{j=1}^{N}(x_j - \overline{x}_N)^2}{Nn}.$$

2.1.3 Loi de la moyenne empirique

Nous donnons ici deux cas, i.e. deux choix de P, où la loi de $\overline{X}_n$ est connue, le cas Gaussien et le cas Bernoulli. Le cas de population finie E est difficile à traiter.

Cas Gaussien On suppose ici que $P = \mathcal{N}_q(\mu, \Sigma^2)$ (voir définition p.26 dans les rappels) avec $\mu \in \mathbb{R}^q$ et Σ^2 une matrice symétrique définie positive de taille $q \times q$. Alors l'échantillon $(X_1, \ldots, X_n)$ suit une loi normale $((\mu, \ldots, \mu)^T, \Sigma_n^2)$ où Σ_n^2 est la matrice $nq \times nq$ de la forme

$$\Sigma_n^2 = \cdots \begin{pmatrix} \Sigma^2 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \Sigma^2 \end{pmatrix}.$$

Par stabilité de la loi normale par transformation affine, on sait que $\overline{X}_n$, qui est bien une transformation affine de l'échantillon $(X_1, \ldots, X_n)$, suit aussi une loi normale

$$\overline{X}_n \sim \mathcal{N}_q(\mathbb{E}(\overline{X}_n), \mathrm{Var}\,(\overline{X}_n)) = \mathcal{N}_q(\mu, n^{-1}\Sigma^2).$$

Cas Bernoulli On suppose ici que $P = \mathcal{B}(p)$ avec $0 < p < 1$. Alors on a

$$n\overline{X}_n \sim \mathcal{B}(n, p)$$

par indépendance des X_i et par définition de la loi Binomiale.

2.1.4 La loi asymptotique de la moyenne empirique

Dans le cadre d'une expérience renouvelable, on peut idéalement faire appel à l'asymptotique et, en utilisant le TLC, on obtient directement sous la condition que X soit de carré intégrable $\mathbb{E}(\|X\|^2) < \infty$:

$$\sqrt{n}\Sigma^{-1}(\overline{X}_n - \mu) \xrightarrow{\mathcal{L}} \mathcal{N}_q(0_q, I_q).$$

Pour n suffisamment grand ($n \geq 100$ en général) on utilise l'approximation normale :

$$\overline{X}_n \overset{\mathcal{L}}{\approx} \mathcal{N}_q(\mu, n^{-1}\Sigma^2).$$

2.2 La matrice de variance-covariance empirique

La matrice de variance-covariance empirique est donnée par l'expression

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

En particulier dans le cas $\mathcal{X} = \mathbb{R}^2$, i.e. $X = (X^{(1)}, X^{(2)})$, on a

$$S_n^2 = \begin{pmatrix} S_n^2(X^{(1)}) & q_{X^{(1)},X^{(2)}} \\ q_{X^{(1)},X^{(2)}} & S_n^2(X^{(2)}) \end{pmatrix}$$

où, pour tout $Y_1, \ldots, Y_n \sim P$ et $Z_1, \ldots, Z_n \sim P'$ dans $\mathbb{R}$ on a la notation

$$S_n^2(Y) = \frac{1}{n}\sum_{j=1}^{n}(Y_j - \overline{Y}_n)^2 \quad \text{et} \quad q_{Y,Z} = \frac{1}{n}\sum_{j=1}^{n}(Y_j - \overline{Y}_n)(Z_j - \overline{Z}_n).$$

On appelle $q_{Y,Z}$ la covariance empirique entre X et Y (rappelons que $\mathrm{Cov}\,(Y,Z) = \mathbb{E}((Y - \mathbb{E}(Y))(Z - \mathbb{E}(Z)))$).

L'e.a. S_n^2 est une matrice aléatoire de taille $q \times q$, de nature plus complexe que le vecteur aléatoire $\overline{X}_n$. Nous allons commencer par étudier son espérance, puis sa variance dans le cas réel $q = 1$ avant d'en déduire sa loi (uniquement dans le cas normal).

2.2.1 L'espérance de S_n^2

Proposition 2.2.1 *Si X est de carré intégrable, alors*

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

Démonstration : Montrons qu'une variante de la formule de Huygens donne la décomposition de S_n^2 suivante

$$S_n^2 = \frac{1}{n}\sum_{j=1}^{n}(X_j - \mu)(X_j - \mu)^T - (\overline{X}_n - \mu)(\overline{X}_n - \mu)^T \tag{2.1}$$

où μ est la vraie moyenne. En effet sachant $(X_1, \ldots, X_n)$ soit l'e.a. discret Y uniformément distribuée sur $\{X_1 - \mu, \ldots, X_n - \mu\}$, i.e. telle que $\mathbb{P}(Y = X_i - \mu) = n^{-1}$. Alors le calcul donne $\mathbb{E}(Y) = \overline{X}_n - \mu$ et $\mathrm{Var}\,(Y) = \mathbb{E}(Y - \mathbb{E}(Y))^2 = S_n^2$ et la formule de Huygens appliquée à Y donne le résultat souhaité.
D'après la décomposition (2.1) on a

$$\mathbb{E}(S_n^2) = \Sigma^2 - \mathrm{Var}\,(\overline{X}_n).$$

La variance de la moyenne empirique vaut $n^{-1}\Sigma_n^2$ d'où le résultat. □

Remarque 3 *L'espérance de la variance empirique n'est pas égale à la vraie variance Σ^2. La matrice de variance-covariance empirique corrigée $S_n^{2\prime}$ est définit par*

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

et vérifie bien que $\mathbb{E}(S_n^{2\prime}) = \Sigma^2$.

2.2.2 La variance de S_n^2 lorsque $q = 1$

Nous ne traitons pas ici la notion de "variance "pour les matrices aléatoires telles que S_n^2. On se restreint au cas réel $\mathcal{X} = \mathbb{R}$; la variance de S_n^2 est donnée par la proposition suivante

Proposition 2.2.2 *Si X^4 est intégrable, i.e. $\mathbb{E}(X^4) < \infty$, alors*

$$Var\,(S_n^2) = \frac{n-1}{n^3}\left((n-1)\mu_4 - (n-3)\sigma^4\right)$$

où $\mu_4 = \mathbb{E}((X - \mu)^4)$ est appelé moment centré d'ordre 4 et $\sigma^4 = Var\,(X)^2$.

Démonstration : Rappelons d'abord que d'après la décomposition (2.1) on a

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

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

$$\begin{aligned}
\mathrm{Var}\,(S_n^2) &= \frac{1}{n}\mathrm{Var}\,(Y_1) - \frac{2}{n}\sum_{j=1}^{n}\mathrm{Cov}\,(Y_j, (\overline{X}_n - \mu)^2) + \mathrm{Var}\,((\overline{X}_n - \mu)^2) \\
&= \frac{1}{n}\mathrm{Var}\,(Y_1) - 2\mathrm{Cov}\,(Y_1, (\overline{X}_n - \mu)^2) + \mathrm{Var}\,((\overline{X}_n - \mu)^2) \\
&= u_n - 2v_n + w_n. \text{ On a d'abord}
\end{aligned}$$

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

D'autre part,

$$\begin{aligned}
v_n &= \mathbb{E}[(X_1 - \mu)^2(\overline{X}_n - \mu)^2] - \mathbb{E}[(X_1 - \mu)^2]\mathbb{E}[(\overline{X}_n - \mu)^2] \\
&= \mathbb{E}[(X_1 - \mu)^2(\overline{X}_n - \mu)^2] - \frac{\sigma^4}{n} \quad \text{avec}
\end{aligned}$$

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

d'où

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

Enfin

$$III_n = Var((\overline{X}_n - \mu)^2) = \mathbb{E}[(\overline{X}_n - \mu)^4] - \frac{\sigma^4}{n^2} \quad \text{où}$$

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

Il s'ensuit que

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

Remarque 4 *Au premier ordre,*

$$Var(S_n^2) \approx \frac{\mu_4-\sigma^4}{n} \qquad \textit{lorsque} \qquad n\to\infty.$$

2.2.3 La loi de S_n^2

Du fait de la complexité de S_n^2 comparativement à $\overline{X}_n$, seul le cas Gaussien réel ($\mathcal{X}=\mathbb{R}$) est envisageable. Supposons donc que $X_1,\ldots,X_n \sim \mathcal{N}(\mu,\sigma^2)$. On a alors la proposition suivante

Théorème 2.2.1 *Dans le cas Gaussien réel, la loi de la variance empirique est déterminée par la formule :*

$$\frac{n}{\sigma^2}S_n^2 \sim \chi^2_{n-1}$$

Démonstration : La démonstration repose sur l'application du Théorème de Cochran (c.f. p. 27 du fascicule "Rappels utiles au cours de statistique mathématique") sur le vecteur Gaussien isotrope $(X_1-\mu,\ldots,X_n-\mu)$ et sur un s.e.v. E de $\mathbb{R}^n$ bien choisi. Étant donné que les X_i sont iid de loi $\mathcal{N}(\mu,\sigma^2)$, on vérifie bien que $\mathbf{X}=(X_1-\mu,\ldots,X_n-\mu)$ est un vecteur Gaussien de $\mathbb{R}^n$ et de loi $\mathcal{N}(0_n,\sigma^2 I_n)$. C'est donc bien un vecteur Gaussien isotrope. On s'intéresse la transformation affine qui à $\mathbf{X}$ associe $\mathbf{X}'=(X_1-\overline{X}_n,\ldots,X_n-\overline{X}_n)$. Si on écrit 1_n la matrice de dimension $n\times n$ qui ne contient que des coefficients 1, alors on vérifie que $\mathbf{X}'=(I_n-n^{-1}1_n)\mathbf{X}$.

Pour prouver que cette transformation $T=I_n-n^{-1}1_n$ est bien une projection π_E on vérifie que $T^2=(I_n-n^{-1}1_n)^2$ est bien égal à T lui-même en utilisant les relations élémentaires $I_n^2=I_n$, $1_nI_n=I_n1_n$ et $1_n^2=n1_n$. On en déduit que ses valeurs propres sont soit égales à 0 soit égales à 1 et donc que le rang de T est la somme de ses valeurs propres, égal à sa trace la somme de ses éléments diagonaux. Ainsi

$$\mathrm{Rg}(T)=\mathrm{Tr}(T)=\frac{n-1}{n}+\cdots+\frac{n-1}{n}=n-1.$$

On en déduit que la dimension du s.e.v. E tel que $T = P_E$ vaut $n-1$. On peut alors appliquer le Théorème de Cochrane et on trouve directement le résultat souhaité :

$$\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 = \|\mathbf{X}'\|^2 = \|P_E(\mathbf{X})\|^2 \sim \chi^2_{n-1}.\square$$

Remarque 5 *Ce résultat est cohérent avec le calcul de la variance de la variance empirique. En effet, on sait que* $\mathbb{E}(Y) = n-1$ *et* $Var(Y) = 2(n-1)$ *pour* $Y \sim \chi^2_{n-1}$ *et donc*

$$Var(S_n^2) = \frac{2\sigma^4(n-1)}{n^2}.$$

On vérifie bien la relation précédente

$$Var(S_n^2) = \frac{n-1}{n^3}\left((n-1)\mu_4 - (n-3)\sigma^4\right)$$

car dans le cas d'un loi normale on a $\mu_4 = 3\sigma^4$. *Cette relation vient du calcul du moment d'ordre 4 d'une loi normale centrée réduite (par IPP) qui donne* $\mathbb{E}(N^4) = 3$, *puis on centre et on réduit la variable* $X \sim \mathcal{N}(\mu, \sigma^2)$ *:*

$$\frac{X-\mu}{\sigma} \overset{\mathcal{L}}{=} N$$

et en prenant le moment d'ordre 4 de cette variable on a

$$\frac{\mu_4}{\sigma^4}\mathbb{E}\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] = \mathbb{E}(N^4) = 3$$

2.2.4 La loi asymptotique de S_n^2

Comme dans le cas de la moyenne empirique, le recours au cadre asymptotique (idéal) permet de donner une approximation normale simple pourvu que n est suffisamment grand (en général $n \geq 100$). Pour simplifier on se restreint au cas $\mathcal{X} = \mathbb{R}$, on a alors le résultat asymptotique :

Théorème 2.2.2 *Soit* $X_1, \ldots, X_n \sim P$ *avec* P *telle que* $\mathbb{E}(|X|^4) < \infty$ *alors on a*

$$\sqrt{n}(S_n^2 - \sigma^2) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \mu_4 - \sigma^4)$$

avec $\mu_4 = \mathbb{E}(X-\mu)^4$.

Démonstration : On commence par appliquer le TLC aux vecteurs $(X_i - \mu)^2$ iid pour tout $1 \leq i \leq n$, d'espérance σ^2 et de variance $\text{Var}\,(X_i-\mu)^2 = \mathbb{E}(X_i-\mu)^4 -$

$(\mathbb{E}(X_i-\mu)^2)^2$ d'après la formule de Huygens, soit $\mathrm{Var}\,(X_i-\mu)^2 = \mu_4-\sigma^4$. On obtient donc

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2-\sigma^2\right)\xrightarrow{\mathcal{L}}\mathcal{N}(0,\mu_4-\sigma^4).$$

D'après la décomposition (2.1) de $S_n^2 = \overline{(X-\mu)^2}_n - (\overline{X}_n-\mu)^2$ on a

$$\sqrt{n}(S_n^2-\sigma^2) = \sqrt{n}(\overline{(X-\mu)^2}_n-(\overline{X}_n-\mu)^2-\sigma^2) = \sqrt{n}(\overline{(X-\mu)^2}_n-\sigma^2)-\sqrt{n}(\overline{X}_n-\mu)^2.$$

Reste à prouver que le dernier terme est négligeable. On sait par le TLC classique que

$$\sqrt{n}(\overline{X}_n-\mu)\xrightarrow{\mathcal{L}}\mathcal{N}(0,\sigma^2).$$

On conclut en utilisant Slutsky avec $\overline{X}_n\xrightarrow{p.s.}\mu \Longrightarrow \overline{X}_n\xrightarrow{P}\mu$ que $\sqrt{n}(\overline{X}_n-\mu)^2\xrightarrow{P}0$ ce qui est suffisant pour prouver le résultat. □

On déduit de cette convergence en loi l'approximation normale

$$S_n^2\overset{\mathcal{L}}{\approx}\mathcal{N}_q(\sigma^2,n^{-1}(\mu_4-\sigma^4))$$

valable pour n grand ($n\geq 100$ en général).

2.3 Le couple $(\overline{X}_n, S_n^2)$

La moyenne et la variance empirique jouent un rôle primordiale en statistique. Nous étudions ici les propriétés du couple $(\overline{X}_n, S_n^2)$ dans le cas $\mathcal{X}=\mathbb{R}$.

2.3.1 L'espérance de $(\overline{X}_n, S_n^2)$

Par définition de l'espérance d'un couple, on trouve simplement $\mathbb{E}(\overline{X}_n, S_n^2) = (\mu, n/(n-1)\sigma^2)$.

2.3.2 La loi de $(\overline{X}_n, S_n^2)$

Dans le cas Gaussien $X_1,\ldots,X_n\sim\mathcal{N}(\mu,\sigma^2)$, on admet le résultat suivant indispensable pour déterminer la densité (et donc la loi) du couple $(\overline{X}_n, S_n^2)$:

Théorème 2.3.1 (Fisher) *Dans le cas Gaussien, $\overline{X}_n$ et S_n^2 sont des v.a. indépendantes.*

On en déduit que la densité du couple et le produit des densités de $\overline{X}_n$ (densité d'une loi normale $\mathcal{N}(\mu,\sigma^2/n)$) et de S_n^2 (densité d'une loi gamma $\gamma((n-1)/2, n/(2\sigma^2))$) soit

$$f_{(\overline{X}_n,S_n^2)}(x,y) = \frac{1}{\Gamma\left(\frac{n-1}{2}\right)\sqrt{2\pi}}\left(\frac{n}{2\sigma^2}\right)^{n/2}y^{(n-3)/2}\exp\left(-\frac{n}{2\sigma^2}((x-\mu)^2+y)\right)1_{y>0}.$$

Remarque 6 *Hors cas Gaussien $\overline{X}_n$ et S_n^2 ne sont pas nécessairement des v.a. indépendantes.*

2.3.3 La loi asymptotique de $(\overline{X}_n, S_n^2)$

En faisant appel au cadre asymptotique on simplifie le problème et on peut déterminer la loi (asymptotique) du couple $(\overline{X}_n, S_n^2)$ pour un grand nombre de lois P dont l'échantillon est issu, i.e. $X_1, \ldots, X_n \sim P$.

Théorème 2.3.2 *Si P est telle que $\mathbb{E}(|X|^4) < \infty$ alors on a*

$$\sqrt{n}\left(\begin{pmatrix}\overline{X}_n \\ S_n^2\end{pmatrix} - \begin{pmatrix}\mu \\ \sigma^2\end{pmatrix}\right) \xrightarrow{\mathcal{L}} \mathcal{N}_2\left(0_2, \begin{pmatrix}\sigma^2 & \mu_3 \\ \mu_3 & \mu_4 - \sigma^4\end{pmatrix}\right)$$

avec $\mu_3 = \mathbb{E}(X-\mu)^3$.

Remarque 7 *Ce n'est pas parce qu'on a prouvé un "TLC" sur $\overline{X}_n$ et sur S_n^2 séparément qu'un "TLC" a forcément lieu sur le couple $(\overline{X}_n, S_n^2)$.*

Démonstration : On applique le TLC classique sur $(X_i, (X_i-\mu)^2)_{i\geq 0}$ une suite iid de vecteurs aléatoires (bien que X_i et $(X_i-\mu)^2$ ne soient pas iid). Comme $\mathbb{E}((X_i, (X_i-\mu)^2)) = (\mu, \sigma^2)$ et de matrice de variance covariance (finie)

$$\Sigma^2 = \begin{pmatrix}\mathrm{Var}_\theta(X) & \mathrm{Cov}_\theta(X, (X-\mu)^2) \\ \mathrm{Cov}_\theta(X, (X-\mu)^2) & \mathrm{Var}_\theta((X-\mu)^2)\end{pmatrix} = \begin{pmatrix}\sigma^2 & \mu_3 \\ \mu_3 & \mu_4 - \sigma^4\end{pmatrix}$$

car $\mathrm{Cov}\,(X, (X-\mu)^2) = \mathbb{E}(X(X-\mu)^2) - \mu\mathbb{E}((X-\mu)^2)$. On obtient alors

$$\sqrt{n}\left(\begin{pmatrix}\overline{X}_n \\ \frac{1}{n}\sum_{i=1}^n (X_i-\mu)^2\end{pmatrix} - \begin{pmatrix}\mu \\ \sigma^2\end{pmatrix}\right) \xrightarrow{\mathcal{L}} \mathcal{N}_2(0_2, \Sigma^2).$$

On conclut comme dans la preuve du théorème sur la loi asymptotique de la variance empirique, à savoir un utilisant la décomposition (2.1) de S_n^2 et que $\sqrt{n}(\overline{X}_n - \mu)^2 \xrightarrow{p.s.} 0$. □

2.4 Moments empiriques

Dans le cas $\mathcal{X} = \mathbb{R}$ il est possible de généraliser les notions de moyenne et de variance empiriques, ce qui donne lieu à la notion des moments empiriques.

Définition 2.4.1 *Soient $X_1, \ldots, X_n \sim P$ et $r \in \mathbb{N}^*$, alors le moment d'ordre r vaut $\mathbb{E}(X^r)$ et sont notés m_r. Le moment centrés d'ordre r vaut $\mathbb{E}((X-\mu)^r)$ où $\mu = m_1$ est la vraie moyenne. Ils ont des équivalent empiriques :*

$$M_n^r = \frac{1}{n}\sum_{j=1}^{n} X_j^r \quad et \quad M_n^{r\prime} = \frac{1}{n}\sum_{j=1}^{n}(X_j - \overline{X}_n)^r$$

appelés le moment empirique d'ordre r et le moment empirique centré d'ordre r.

Moyenne et variance en sont des cas particuliers car $\mu = m^1$ et $\sigma^2 = \mu_2$, d'où les versions empiriques $M_n^1 = \overline{X}_n$ et $M_n^{2\prime} = S_n^2$. Le moment centré d'ordre 1 μ_1 vaut toujours 0.

2.4.1 L'espérance du moment empirique

La linéarité de l'espérance nous garantit que $\mathbb{E}(M_n^r) = m_r$. Par contre $\mathbb{E}(M_n^{r\prime}) \neq \mu_r$ et on peut corriger le moment empirique centré (c.f. cas $r = 2$ où $M_n^{2\prime} = S_n^2$).

2.4.2 Loi asymptotique du moment empirique

Une application du TLC nous donne la loi asymptotique des moments (centrés ou non) :

Proposition 2.4.1 *Si $\mathbb{E}(X^{2r}) < +\infty$, i.e. m_{2r} existe, alors*

$$\begin{aligned}\sqrt{n}(M_n^r - m_r) &\xrightarrow{\mathcal{L}} \mathcal{N}(0, m_{2r} - m_r^2)\\ \sqrt{n}(M_n^{r\prime} - \mu_r) &\xrightarrow{\mathcal{L}} \mathcal{N}(0, \mu_{2r} - \mu_r^2).\end{aligned}$$

démonstration : Application directe du TLC à X_i^r pour le cas M_n^r.
Application du TLC à $(X_i - m_1)^r$ dans le cas $M_n^{r\prime}$ puis Slustky en utilisant la LGN $\overline{X}_n \xrightarrow{P} m_1$. □

Dans le cas $M_n^1 = \overline{X}_n$ et $M_n^{2\prime} = S_n^2$ on retrouve les résultats trouvés précédemment.

2.5 Fonction de répartition empiriques

Dans le cas $\mathcal{X} = \mathbb{R}$ la fonction de répartition empirique empirique est définie par :

Définition 2.5.1 *Soit $X_1, \ldots, X_n \sim P$. La fonction de répartition empirique F_n est définie par la fonction qui à $x \in \mathbb{R}$ asocie*

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

On a donc $F_n : \mathbb{R} \to [0; 1]$ qui est croissante, continue à droite et admettant une limite à gauche (cadlag) par définition.

Remarquons que F_n est la fonction de répartition de la loi $\mathcal{U}(\{X_1, \ldots, X_n\})$. F_n est donc une fonction aléatoire dont l'étude de la loi dépasse le cadre de ce cours. On se restreint donc à l'étude de la loi de $F_n(x)$ avec $x \in \mathbb{R}$ fixé qui est une variable aléatoire.

2.5.1 La loi de $F_n(x)$ avec $x \in \mathbb{R}$ fixé

Pour tout $x \in \mathbb{R}$ fixé, on effectue le changement de variable aléatoire en considérant $Y_i = 1_{X_i \leq x}$. Il est facile de voir que $F_n(x) = \overline{Y}_n$ et que $Y = 1_{X \leq x}$ est une variable aléatoire valant soit 0 soit 1, donc $Y \sim \mathcal{B}(p)$ avec $p = \mathbb{P}(Y = 1) = \mathbb{E}(Y)$. Ici on trouve facilement $p = \mathbb{E}(1_{X \leq x}) = \mathbb{P}(X \leq x) = F(x)$. on en déduit, d'après l'étude de la loi de la moyenne empirique dans le cas Bernoulli que $nF_n(x) \sim \mathcal{B}(n, F(x))$. De plus,

$$\mathbb{E}(F_n(x)) = \mathbb{E}(\overline{Y}_n) = \mathbb{E}(Y) = F(x) \text{ et}$$
$$\text{Var}\,(F_n(x)) = \text{Var}\,(F_n(x)) = n^{-1}\text{Var}\,(Y) = \frac{F(x)(1 - F(x))}{n}.$$

2.5.2 La loi asymptotique de $F_n(x)$ avec $x \in \mathbb{R}$ fixé

En appliquant le TLC aux $Y_i = 1_{X_i \leq x}$, on trouve :

Théorème 2.5.1 *Soit $F = F_X$ la fonction de répartition de X alors $\forall\, x \in \mathbb{R}$*

$$\sqrt{n}(\mathbb{F}_n(x) - F(x)) \xrightarrow{\mathcal{L}} \mathcal{N}(0, F(x)(1 - F(x)).$$

Remarque 8 *En tant que fonction, $F_n : x \mapsto F_n(x)$ est une fonction aléatoire constante par morceau en x qui a des sauts de hauteur n^{-1} en chacun de ses points de discontinuité $(X_1, \ldots, X_n)$ (le saut peut être double en un point $X_i = X_j$ pour $j \neq i$). La "densité empirique" est sa "dérivée au sens des distribution", i.e. la mesure empirique $P_n = n^{-1} \sum_{i=1}^{n} \delta_{X_i}$.*

Chapitre 3

Théorie de l'information de Fisher

Soit $X_1 \ldots, X_p \sim P$ un échantillon d'observations iid à valeurs dans $\mathcal{X}$ ($\mathbb{R}^q$ pour $q \geq 1$) muni de sa tribu des Boréliens $\mathcal{B}$. Pour inférer sur les caractéristiques d'une propriété X inconnue, le statisticien utilise des fonctions de l'échantillon : $T_n = T(X_1, \ldots, X_n)$. Ces éléments aléatoires sont appelés des statistiques. La statistique T_n doit contenir autant d'information que l'échantillon $(X_1, \ldots, X_n)$ pour l'inférence du caractère inconnu. La théorie de l'information de Fisher fournit un cadre mathématique pour quantifier l'information contenue par l'échantillon $(X_1, \ldots, X_n)$. Mais commençons par étudier les différentes propriétés de la statistique T_n.

3.1 Propriétés des statistiques

3.1.1 Définition de la statistique

Soit $(\mathcal{Y}, \mathcal{C})$ l'espace des caractéristiques que l'on souhaite inférer sur la loi P à partir des observations $(X_1, \ldots, X_n)$.

Définition 3.1.1 *Soit T une application mesurable de $\mathcal{X}^n$ dans $\mathcal{Y}$ ne dépendant pas des caractéristiques inconnues de la loi P dont est issu l'échantillon, i.e. $T : \mathcal{X}^n \to \mathcal{Y}$. Alors $T_n = T(X_1, \ldots, X_n)$ est un élément aléatoire de $\mathcal{Y}$ appelé statistique.*

Exemple 3.1.1
- *Toutes les caractéristiques empiriques $T_n = \varphi(P_n)$ sont des statistiques.*
- *La moyenne empirique $\overline{X}_n$ est une statistique, i.e. peut s'écrire $T(X_1, \ldots, X_n)$ avec $T : \mathcal{X}^n \to \mathcal{Y} = \mathcal{X}$,*
- *Pour $d \in \mathbb{N}$ tel que $\mathcal{X} = \mathbb{R}^q$ la variance empirique S_n^2 est une statistique, i.e. peut s'écrire $T(X_1, \ldots, X_n)$ avec $T : \mathcal{X} \to \mathcal{Y} = \mathcal{S}_d^+(\mathbb{R})$ où $S_d^+(\mathbb{R})$ est l'espace des matrices symétriques positives,*

- *Pour $\mathcal{X} = \mathbb{R}$ les moments empiriques (centrés) d'ordres quelconques sont tous des statistiques à valeurs dans $\mathcal{Y} = \mathbb{R}$,*
- *Pour $\mathcal{X} = \mathbb{R}$ la fonction de répartition empirique F_n est une statistique à valeur dans $\mathcal{Y} = \mathcal{D}$, l'ensemble des fonctions càdlàg de $\mathbb{R}$ dans $\mathbb{R}$.*

Exemple 3.1.2 (Statistique de Student) *Pour $\mathcal{X} = \mathbb{R}$ et $X_1, \cdots, X_n \sim P$ la statistique de Student est la quantité*

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

Si $P = \mathcal{N}(\mu, \sigma^2)$ alors $T(\mu)$ suit une loi de Student à $(n-1)$ degrés de liberté (ce résultat découle directement de la définition de la loi de Students et du théorème de Fisher).

3.1.2 Statistique d'ordre

Dans le cas $\mathcal{X} = \mathbb{R}$, certaines statistiques ne dépendent de l'échantillon que lorsque celui-ci est ordonné en ordre croissant :

Définition 3.1.2 *L'<u>échantillon ordonné</u> dans l'ordre croissant, noté $(X_{(1)}, \ldots, X_{(n)})$, est défini tel que $X_{(k)}$ soit la k-ème plus petite valeur de l'échantillon $(X_1, \ldots, X_n)$. Alors $T_n = T(X_{(1)}, \ldots, X_{(n)})$ est une <u>statistique d'ordre</u>.*

Exemple 3.1.3

- *Pour tout j, $X_{(j)}$ est une statistiques d'ordre appelée statistique d'ordre de rang j.*
- *$X_{(1)} = \min(X_1, \ldots, X_n)$ et $X_{(n)} = \max(X_1, \ldots, X_n)$ sont des statistiques d'ordre.*

Contrairement à l'échantillon $(X_1, \ldots, X_n)$, l'échantillon ordonné n'est pas iid. Dans le cas où X est absolument continue (admet une densité notée f), on peut toutefois spécifier la loi de l'échantillon ordonné :

Théorème 3.1.1 *Le vecteur ordonné $(X_{(1)}, \ldots, X_{(n)})$ a pour densité*

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

Démonstration : Soit σ une permutation aléatoire suivant la loi $\mathcal{U}(\{\text{permutations de } \{1, \ldots, n\}\})$ indépendante de $(X_1, \ldots, X_n)$. Par indépendance, on obtient

$$\mathbb{P}((X_{\sigma(1)}, \ldots, X_{\sigma(1)}) \in]x_{\sigma(1)} - h, x_{\sigma(1)}] \times \cdots \times]x_{\sigma(n)} - h, x_{\sigma(n)}])$$
$$\frac{1}{n!} \sum_{permutations} \prod_{i=1}^{n} (F(x_i) - F(x_i - h)) = \prod_{i=1}^{n} (F(x_i) - F(x_i - h)).$$

D'autre part

$$\begin{aligned}&\mathbb{P}((X_{\sigma(1)},\dots,X_{\sigma(1)}) \in]x_{\sigma(1)}-h, x_{\sigma(1)}] \times \cdots \times]x_{\sigma(n)}-h, x_{\sigma(n)}])\\ &= \mathbb{P}((X_{(1)},\dots,X_{(n)}) \in]x_{(1)}-h, x_{(1)}] \times \cdots \times]x_{(n)}-h, x_{(n)}]) \mid \sigma(\cdot)=(\cdot))\mathbb{P}(\sigma(\cdot)=(\cdot)).\end{aligned}$$

Comme la loi de $(X_1,\dots,X_n)$ est elle aussi absolument continue, on se restreint au cas $x_i \neq x_j$ pour $i \neq j$. Il existe alors une unique permutation σ' telle que $x_{\sigma'(1)} < \ldots < x_{\sigma'(n)}$. Pour h suffisamment petit, $\sigma'(\cdot)=(\cdot)$ p.s.. Cette permutation σ' ne dépend que de $(x_1,\dots,x_n)$, elle est indépendante de $(X_1,\dots,X_n)$. De même, pour h suffisamment petit, l'événement $\{\sigma(\cdot)=(\cdot)\}=\{\sigma=\sigma'\}$ est indépendant de $(X_1,\dots,X_n)$ car σ l'est par définition. On obtient

$$\begin{aligned}&\mathbb{P}((X_{\sigma(1)},\dots,X_{\sigma(n)}) \in]x_{\sigma(1)}-h, x_{\sigma(1)}] \times \cdots \times]x_{\sigma(n)}-h, x_{\sigma(n)}])\\ &\qquad = \mathbb{P}((X_{(1)},\dots,X_{(n)}) \in]x_{(1)}-h, x_{(1)}] \times \cdots \times]x_{(n)}-h, x_{(n)}]))\mathbb{P}(\sigma=\sigma').\end{aligned}$$

Par définition de la loi uniforme, $\mathbb{P}(\sigma=\sigma')=1/n!$ et en posant $z_i=x_{(i)}$ le résultat est prouvé. □

On en déduit les lois de chaque marginale de l'échantillon ordonnée $X_{(k)}$. Remarquons que les $X_{(k)}$ ne sont pas identiquement distribués, leurs densité dépend de k :

Théorème 3.1.2 *Si F est la fonction de répartition de X, alors a statistique d'ordre $X_{(k)}$ a pour densité*

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

De plus, $X_{(i)}$ dépend de $X_{(j)}$ pour $i \neq j$, leur densité jointe n'est pas le produit de leurs densités :

Théorème 3.1.3 *La loi jointe du couple $(X_{(i)},X_{(j)}), i<j$ admet pour densité*

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

3.1.3 Statistique paramétrique

Soit une statistique $T_n \in \mathcal{Y}$, deux cas sont possibles :
- Soit $\mathcal{Y}$ est un ensemble inclus dans un espace de dimension fini : il existe $d \in \mathbb{N}$ pour lequel $\mathcal{Y} \subseteq \mathbb{R}^d$,

– Soit $\mathcal{Y}$ n'est inclus dans aucun espace de dimension fini.

Dans le premier cas $T_n \in \mathbb{R}^d$ est appelée une statistique paramétrique de dimension d. Dans le second cas la statistique T_n est de dimension infinie ; c'est une statistique non paramétrique.

Définition 3.1.3 *On appelle modèle paramétrique de paramètre $\theta \in \Theta$ pour un certain espace de dimension fini $\Theta \subseteq \mathbb{R}^d$, $d \geq 1$, le couple (P_θ, Θ), où P_θ est la loi de probabilité de X qui dépend du paramètre θ inconnu et Θ est l'ensemble des paramètres θ envisageables.*

On notera simplement $X_1, \ldots, X_n \sim P_\theta$ l'échantillon issue du modèle paramétrique (P_θ, Θ) en spécifiant bien l'espace des paramètres Θ.

Exemple 3.1.4 *Cas de l'expérience succès-echec : $X_1, \ldots, X_n \sim \mathcal{B}(\theta)$ avec $0 < \theta < 1$ veut dire qu'on se place dans le cadre d'un échantillon issu de l'expérience X qui a pour valeur 0 ou 1 (loi de Bernoulli), que cette loi dépend uniquement de la probabilité de succès ($X = 1$) notée θ, que cette caractéristique est inconnue est qu'on recherche à inférer dessus à partir de l'échantillon $(X_1, \ldots, X_n)$. Dans ce cas $\Theta =]0; 1[$.*

Exemple 3.1.5

- *les statistiques $\overline{X}_n$, S_n^2, M_n^r, $M_n^{r'}$ et $F_n(x)$ avec $x \in \mathbb{R}$ fixé sont des statistiques paramétriques,*
- *la statistique F_n est non-paramétrique.*

Dans toute la suite de ce cours on se limitera au cadre d'un échantillon $(X_1, \ldots, X_n)$ issu d'un modèle paramétrique, où la caractéristique à inférer est le paramètre θ.

3.1.4 Statistique exhaustive et statistique libre

Soit le modèle paramétrique $X_1, \ldots, X_n \sim P_\theta$ avec $\theta \in \Theta$.

Définition 3.1.4 *La statistique T_n sera dite exhaustive pour θ si la loi conditionnelle de l'échantillon $(X_1, \ldots, X_n)$ sachant $T_n = t$ n'est pas une fonction du paramètre θ :*

$$P_\theta((X_1, \ldots, X_n) \in \cdot \mid T_n = t) \quad \textit{ne dépend pas de } \theta.$$

Remarque 9 *Lorsque la valeur prise par la statistique exhaustive T_n est connue (égale à t), alors l'échantillon $(X_1, \ldots, X_n)$ ne fournit plus d'information sur le paramètre inconnu θ car sa loi ne dépend plus de θ. La statistique exhaustive contient toute l'information nécessaire à l'inférence de θ.*

On notera $f(x,\theta)$ la densité de P_θ relativement à une mesure dominante et $\sigma-$finie, ν. On va se restreindre au cas où ν est la mesure de Lebesgue (variables aléatoires de loi absolument continue) et on retrouve la densité f notée f_θ ou la mesure de comptage (variables aléatoires de loi discrète) et on retrouve le système $P_\theta(X=x)$. L'indice θ est ajouté aux notation usuelles f et P pour faire remarquer que la loi des observations X_i dépend de ce paramètre inconnu. L'existence d'une densité permet de trouver facilement une statistique exhaustive grâce au théorème suivant :

Théorème 3.1.4 ***(Théorème de factorisation)*** *Soit T une fonction mesurable de $(\mathcal{X}^n,\mathcal{B}_n)\to(\mathcal{Y},\mathcal{C})$. Alors $T_n=T(X_1,\ldots,X_n)$ est une statistique exhaustive pour θ si et seulement s'il existe deux fonctions mesurables $g:\mathcal{Y}\times\Theta\to\mathbb{R}^+$ et $h:\mathcal{X}^n\to\mathbb{R}^+$ telles que la densité $f(x_1,\ldots,x_n;\theta)$ de l'échantillon $(X_1,\ldots,X_n)$ se mette sous la forme*

$$f(x_1,\ldots,x_n;\theta)=h(x_1,\ldots,x_n)g(T(x_1,\ldots,x_n),\theta).$$

Démonstration : On ne montre que l'implication "densité factorisée" implique "identification d'une statistique exhaustive".

Soit ℓ la densité conditionnelle de $(X_1,\ldots,X_n)$ sachant que $T(X_1,\ldots,X_n)=t$. Soit $T^{-1}(t)=\{(x_1,\ldots,x_n)\in\mathbb{R}^n\ /\ T(x_1,\ldots,x_n)=t\}$. On peut alors écrire

$$\begin{aligned}\ell(x_1,\ldots,x_n) &= \frac{f(x_1,\ldots,x_n;\theta)1_{T(x_1,\ldots,x_n)=t}}{\int_{T^{-1}(t)}f(x_1,\ldots,x_n;\theta)d\nu^{\otimes n}(x_1,\ldots,x_n)}\\ &= \frac{h(x_1,\ldots,x_n)g(t,\theta)}{g(t,\theta)\int_{T^{-1}(t)}h(x_1,\ldots,x_n)d\nu^{\otimes n}(x_1,\ldots,x_n)}\\ &= \frac{h(x_1,\ldots,x_n)}{\int_{T^{-1}(t)}h(x_1,\ldots,x_n)d\nu^{\otimes n}(x_1,\ldots,x_n)}.\end{aligned}$$

La fonction ℓ ne dépend plus de θ donc le résultat est prouvé. □

Exemple 3.1.6

– *Soit $X_1,\ldots,X_n\sim\mathcal{U}[0,\theta]$. On a*

$$f(x_1,\ldots,x_n;\theta)=\frac{1}{\theta^n}1_{0\le\min_{1\le i\le n}x_i}1_{\sup_{1\le i\le n}x_i\le\theta}.$$

En posant

$$h(\mathbf{x})=1_{0\le\min_{1\le i\le n}x_i}\quad et\quad g(T(x_1,\ldots,x_n),\theta)=\frac{1}{\theta^n}1_{T(x_1,\ldots,x_n)\le\theta}$$

on déduit que $T(X_1,\ldots,X_n)=\max_{1\le j\le n}X_j=X_{(n)}$ est une statistique (d'ordre) exhaustive pour θ.

– *Soit* $X_, \ldots, X_n \sim \mathcal{E}(\theta)$. *On a*

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

et donc

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

est bien une statistique exhaustive pour θ.

– *Soit* $X_1, \ldots, X_n \sim \mathcal{P}(\theta)$. *On a*

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

et donc

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

est bien une statistique exhaustive pour θ.

– *Soit* $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$. *Alors la statistique*

$$T(X_1, \ldots, X_n) = \left(\frac{1}{n}\sum_{j=1}^{n} X_j, \frac{1}{n}\sum_{j=1}^{n} X_j^2\right)$$

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

La statistique libre est l'opposé de la statistique exhaustive : c'est une statistique qui ne contient pas d'information pour l'inférence du paramètre θ.

Définition 3.1.5 *Une statistique* T *d'un modèle paramétrique est dite libre si sa loi ne dépend pas du paramètre* θ.

N'apportant aucune information pour l'estimation du paramètre θ, une statistique libre est ce qu'on appelle un paramètre de nuisance.

3.2 Information au sens de Fisher

On définit dans cette section une quantité mathématique mesurant l'information contenue dans un modèle statistique. On verra dans la section suivante que cette définition d'information due à Fisher concorde avec l'heuristique faite sur les notions de statistique exhaustive et de statistique libre, à savoir la première

contient toute l'information de l'échantillon $(X_1, \ldots, X_n)$ pour inférer sur θ, la seconde ne contient au contraire aucune d'information pour inférer sur θ.

Soit le modèle paramétrique (P_θ, Θ). La définition de l'information de Fisher dépend de la notion de Score. Pour que cette notion soit bien définie, on suppose que les hypothèses suivantes sur la densité $f(x, \theta)$ de P_θ relativement à la mesure dominante ν sont satisfaites. On se place dans le contexte d'un modèle régulier :

Définition 3.2.1 *Soit (P_θ, Θ) un modèle paramétrique. On note $f(x, \theta)$ la densité de P_θ relativement à la mesure dominante ν (mesure de comptage ou mesure de Lebesgue). Le modèle (P_θ, Θ) est régulier si les 4 hypothèses suivantes sont satisfaites :*

(H1) *L'ensemble des paramètres Θ est un ouvert de $\mathbb{R}^d$ pour d fini et*

$$f(x, \theta) > 0 \quad \Longleftrightarrow \quad f(x, \theta') > 0, \qquad \forall \theta, \theta' \in \Theta.$$

(H2) *Pour ν presque tout x, les fonctions $\theta \mapsto f(x, \theta)$ et $\theta \mapsto \log f(x, \theta)$ sont deux fois continûment dérivables sur Θ.*

(H3) *Pour tout $\theta^* \in \Theta$ il existe un ouvert $U_{\theta^*} \subseteq \Theta$ contenant θ^* et une fonction borélienne $\Lambda(x)$ tels que*

$$\|\nabla_\theta(\log f(x, \theta))\| \leq \Lambda(x) \quad et \quad \|\mathbb{H}_\theta(\log f(x, \theta))\| \leq \Lambda(x)$$

pour tout $\theta \in U_{\theta^*}$ et ν-presque tout $x \in \mathcal{X}$, et

$$\int \Lambda(x) \sup_{\theta \in U_{\theta^*}} f(x, \theta) d\nu(x) < \infty.$$

(H4) *La matrice $-\mathbb{E}_\theta[\mathbb{H}_\theta(\log f(X, \theta)]$ de taille $d \times d$ est symétrique définie positive pour tout $\theta \in \Theta$.*

Exemple 3.2.1 *Les modèles de Poisson $(\mathcal{P}(\theta), \theta > 0)$, exponentiel $(\mathcal{E}(\lambda), \lambda > 0)$ et Gaussien $(\mathcal{N}(\mu, \sigma^2), \mathbb{R} \times \mathbb{R}_+^*)$ sont réguliers mais le modèle Uniforme $(\mathcal{U}[0, \theta], \theta > 0)$ ne vérifie pas* **(H1)**.

Supposons par la suite que le modèle paramétrique $(\mathcal{X}, P_\theta)$ soit régulier. Alors on peut définir la notion de vecteur score :

Définition 3.2.2 *On appelle score pour une expérience aléatoire $X \sim P_\theta$ le vecteur aléatoire $S(X, \theta)$ définit par*

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

Propriété 1

- *Le score est un vecteur aléatoire centré*

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

Notons que l'espérance $\mathbb{E}_\theta$ est prise par rapport à P_θ, où θ à la même valeur que dans l'expression $S(X,\theta)$.

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

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

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

A partir du vecteur score on définit facilement l'information de Fisher :

Définition 3.2.3 *L'information de Fisher d'un modèle paramétrique régulier (P_θ, Θ)la fonction qui à toute valeur du paramètre inconnu $\theta \in \Theta \subseteq \mathbb{R}^d$ associe une matrice de taille $d \times d$ $I(\theta)$ vérifiant*

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

On a les propriétés suivantes

Propriété 2 *Par définition, l'information de Fisher est une matrice symétrique définie positive en tant que matrice de variance-covariance du vecteur score (car le vecteur score est centré). Pour tout $1 \le i,j \le d$*

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

Donc $I(\theta) = -\mathbb{E}[\mathbb{H}_\theta(\log f(X,\theta)]$ est une matrice symétrique définie positive sous l'hypothèse **(H4)**.

Exemple 3.2.2 *Soit $X \sim \mathcal{N}(\mu,\sigma^2)$, alors*

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

En effet,

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

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

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

$$\frac{\partial^2 \log f(x,\mu,\sigma^2)}{\partial\mu\partial\sigma^2} = \frac{\mu-x}{\sigma^4} \Rightarrow \mathbb{E}_\theta\left[\frac{\partial^2 \log f(X,\mu,\sigma^2)}{\partial\mu\partial\sigma^2}\right] = 0.$$

Pour l'échantillon $(X_1,\ldots,X_n)$, le vecteur score $S((X_1,\ldots,X_n),\theta)$ sera noté $S_n(\theta)$ et l'information de Fisher associée sera notée $I_n(\theta)$. Par indépendance, on a

$$S_n(\theta) = \nabla_\theta\left(\sum_{i=1}^n \log f(X_i,\theta)\right) = \sum_{j=1}^n S(X_j,\theta).$$

Or les vecteurs scores $S(X_1,\theta),\ldots,S(X_n,\theta)$ sont iid (de même loi que $S(X,\theta)$). On a donc la relation

$$I_n(\theta) = \mathrm{Var}_\theta(S_n(\theta)) = \sum_{j=1}^n \mathrm{Var}_\theta S(X_j,\theta) = nI(\theta).$$

Enfin, remarquons que le TLC appliqué aux $S(X_i,\theta)$ donne immédiatement la loi asymptotique du score. Pour tout $\theta\in\Theta$ on a :

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

3.3 Lien entre l'information au sens de Fisher et la statistique

Le résultat suivant établit le lien étroit qui existe entre les notions de statistique et d'information au sens de Fisher. Il valide la notion d'information choisie par Fisher. Soit $I_n(\theta) = nI(\theta)$ l'information de Fisher de l'échantillon $(X_1,\ldots,X_n)$ issu du modèle paramétrique régulier (P_θ,Θ).

Considérons T_n une statistique $T(X_1,\ldots,X_n)$. Soit $P_\theta^{T_n}$ la loi de la statistique T_n associé à $X_1,\ldots,X_n\sim P_\theta$ et soit $I_{T_n}(\theta)$ l'information contenue dans le modèle régulier $(P_\theta^{T_n},\Theta)$ (on suppose qu'il vérifie aussi (H1)-(H4)). On rappelle que pour deux matrices A et B on a $A\leq B \Leftrightarrow B-A$ est une matrice symétrique positive.

Théorème 3.3.1 *Pour toute statistique T_n on a la relation suivante*

$$I_{T_n}(\theta) \leq I_n(\theta)$$

et

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

Exemple 3.3.1 *Soit $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ et considérons la statistique*

$$T_n = S_n^2 = \frac{1}{n}\sum_{j=1}^{n}(X_i - \overline{X}_n)^2.$$

On sait que

$$\frac{n}{\sigma^2}T_n \sim \chi^2_{n-1}.$$

Etant donné que $\chi^2_{n-1} = \gamma((n-1)/2, 1/2)$, on utilise la stabilité de la loi Gamma pour obtenir $T_n \sim \gamma((n-1)/2, n/(2\sigma^2))$ de densité

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

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

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

où $cste(t)$ est une constante qui ne dépend pas de σ. D'où la dérivée seconde

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

d'où

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

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

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

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

Deuxième partie

L'estimation statistique

Préambule

Soit X un e.a. à valeurs dans $(\mathcal{X}, \mathcal{B})$ avec $\mathcal{X} \subseteq \mathbb{R}^q$ issue du modèle statistique (P_θ, Θ) avec $\Theta \subseteq \mathbb{R}^d$.

Définition 3.3.1 *Le paramètre d'intérêt θ détermine complètement la loi P_θ, i.e. θ est le vecteur composée de tous les paramètres inconnus du statisticien.*

La densité de P_θ par rapport à une mesure dominante σ- finie ν (mesure de comptage dans le cas d'une loi discrète, la mesure de Lebesgue dans le cas d'une loi absolument continue) sera notée $f(x, \theta)$.

Dans le cadre de l'estimation ponctuelle, l'objectif du statisticien est de déterminer la vraie valeur du paramètre θ de la loi P_θ dont est issu l'échantillon $(X_1, \ldots, X_n)$. A partir de l'information fournie par cet échantillon $(X_1, \ldots, X_n)$ le statisticien utilise des statistiques $T_n \in \mathcal{Y}$ pour approcher θ, donc $\mathcal{Y} = \Theta$ et T_n ne doit pas dépendre de θ inconnu.

Définition 3.3.2 *Toute statistique $T_n \in \mathcal{Y}$ telle que $\mathcal{Y} = \Theta$ est appelée un estimateur (ponctuel) du paramètre $\theta \in \Theta$.*

Un estimateur $T_n = T(X_1, \ldots, X_n)$ est donc un e.a. de Θ. Une réalisation $T(x_1, \ldots, x_n)$ de T_n sera appelée une estimation de θ et notée $\hat{\theta}_n$. Par abus, la notation $\hat{\theta}_n$ désigne aussi souvent l'estimateur. Nous étudions dans le prochain chapitre l'approche non asymptotique pour une certaine famille de modèle. Puis dans un second chapitre nous traiterons de l'approche asymptotique dans un cadre plus général. Dans un troisième chapitre nous traiterons d'un autre type d'estimation : l'estimation par régions de confiance.

Chapitre 4

Approche non asymptotique

4.1 Critères de comparaison d'estimateurs

Nous allons donner des critères non asymptotiques de la qualité d'un estimateur. Soit T_n et T'_n deux estimateurs de θ. Ce sont des e.a. de $\Theta \subseteq \mathbb{R}^d$ pour $d \geq 1$. On munit $\mathbb{R}^d$ de la norme $\|\cdot\|$ associée au produit scalaire usuel.

4.1.1 Le risque quadratique

On doit donc choisir un critère qui permettra au statisticien de comparer différents estimateurs. Un bon critère est le risque quadratique :

$$R_n(T_n, \theta) = \mathbb{E}_\theta \|T_n - \theta\|^2.$$

On en déduit la définition suivante

Définition 4.1.1 *Si, pour tout $\theta \in \Theta$ on a*

$$R_n(T_n, \theta) \leq R_n(T'_n, \theta),$$

et si il existe un $\theta' \in \Theta$ tel que

$$R_n(T_n, \theta') < R_n(T'_n, \theta'),$$

alors T_n est un meilleur (préférable) estimateur que T'_n et T'_n est un estimateur inadmissible.
Un estimateur T_n est dit admissible si il n'existe pas d'estimateur meilleur que T_n.

L'erreur quadratique moyenne de T_n se décompose en deux termes, le biais et variance de l'estimateur T_n.

4.1.2 Décomposition biais-variance du risque

Définition 4.1.2 *On appelle biais de l'estimateur T_n la quantité $b_\theta(T_n) = \mathbb{E}_\theta(T_n) - \theta$. Un estimateur T_n de θ est dit sans biais ou non-biaisé si*

$$b_\theta(T_n) = 0 \quad soit \quad \mathbb{E}_\theta(T_n) = \theta.$$

Exemple 4.1.1 *Soit le modèle $(P_\theta, \theta = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+^*)$ tel que $\mathbb{E}_\theta(X) = \mu$ et $Var_\theta(X) = \sigma^2$. Alors $T_n = (\overline{X}_n, S_n^{2\prime})^T$ est un estimateur sans biais de (μ, σ^2).*

Remarque 10 *La définition du biais nécessite l'intégrabilité de T_n : $\mathbb{E}_\theta\|T_n\| < \infty$.*

Lorsque de plus T_n est de carré intégrable, i.e. $\mathbb{E}_\theta\|T_n\|^2 < \infty$, on a la décomposition biais-variance du risque quadratique :

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

où $\mathrm{Var}\,(T_n)$ est la matrice variance-covariance de T_n. Cette décomposition permet de se ramener à une discussion sur la variance pour les estimateurs sans biais.

4.1.3 Comparaison des variances des estimateurs sans biais

D'après la décomposition biais-variance, la comparaison d'estimateurs sans biais revient à la comparaison de leurs variances ; on parle alors d'efficacité. Dans cette section, on se limite donc au cas où T_n et T_n' sont deux estimateurs sans biais de θ.

Définition 4.1.3 *L'estimateur T_n est dit plus efficace que T_n' s'il est meilleur au sens de la variance :*

$$Var_\theta(T_n) \leq Var_\theta(T_n'), \quad \forall\, \theta \in \Theta \quad et \quad \exists \theta' \in \Theta, \quad Var_{\theta'}(T_n) < Var_{\theta'}(T_n').$$

On dit que l'estimateur sans biais T_n est de variance minimale si $Var_\theta(T_n) \leq Var_\theta(T_n')$ pour tout estimateur sans biais T_n' et pour tout $\theta \in \Theta$.

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

4.1.4 Modèles réguliers et efficacité d'estimateurs

Dans le cadre d'un modèle régulier, c.f. Définition 3.2.1, l'information de Fisher est bien définie (il n'y a plus de problème d'intégrabilité). De plus, comme toute matrice symétrique définie positive, elle est inversible. Il est alors possible de donner un critère absolu pour les estimateurs de variance minimale en fonction de l'inverse de l'information de Fisher.

Théorème 4.1.1 *Soit $T_n = T(X_1, ..., X_n)$ un estimateur sans biais de $\theta, \theta \in \Theta$ de carré intégrable $\mathbb{E}_\theta \|T_n\|^2 < \infty$. Alors on a*

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

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

Démonstration : On note $S = S((X_1, \dots X_n), \theta) = \nabla_\theta \log f((X_1, \dots, X_n), \theta)$ le vecteur score. On sait que $\mathbb{E}_\theta(S) = 0$ et $\mathrm{Var}_\theta(S) = I_n(\theta)$ pour tout $\theta \in \Theta$. D'autre part, T_n étant un estimateur sans biais de θ, on a $\mathbb{E}_\theta(T_n) = \theta$ donc en dérivant

$$\int_{\mathcal{X}^n} T_n(x_1, \dots, x_n)(\nabla_\theta f((x_1, \dots, x_n), \theta))^T d\nu(x) = I_d.$$

Le vecteur score s'écrit $f((x_1, \dots, x_n), \theta) S((x_1, \dots, x_n), \theta) = \nabla_\theta f((x_1, \dots, x_n), \theta)$ et on obtient $\mathbb{E}_\theta(T_n S^T) = I_d$ $(= \mathbb{E}_\theta(S T_n^T))$.

En utilisant ce qui précède et le fait que $I_n^T = I_n$ pour tout θ, on a

$$\begin{aligned} \mathrm{Var}_\theta(I_n^{-1} S - T_n) &= I_n^{-1} \mathrm{Var}_\theta(S_n) I_n^{-1} - I_n^{-1} \mathbb{E}_\theta(S T_n^T) - \mathbb{E}_\theta(T_n S^T) I_n^{-1} + \mathrm{Var}(T_n) \\ &= \mathrm{Var}(T_n) - I_n^{-1}. \end{aligned}$$

Comme $\mathrm{Var}(T_n) - I_n^{-1}$ s'exprime aussi comme une matrice de variance-covariance (positive), le théorème est prouvé. □

Définition 4.1.4 *Un estimateur sans biais T_n dont la matrice de variance-covariance satisfait l'égalité*

$$Var_\theta(T_n) = I_n^{-1}(\theta)$$

est appelé un estimateur efficace.

Remarque 11

- *Le critère d'efficacité n'a de sens que pour discriminer les estimateur sans biais.*
- *Un estimateur efficace est de variance minimale.*

- *Rien ne garantit l'existence d'un estimateur dont la variance atteint la borne de Cramér-Rao.*
- *Un estimateur peut être sans biais, de variance minimale, mais ne pas atteindre la borne de Cramer-Rao, donc ne pas être efficace.*
- *L'efficacité est une notion qui fait le lien entre la théorie de l'information et l'estimation : plus l'information de Fisher est grande et plus la borne de Cramer Rao est petite, i.e. plus on a une chance de trouver un estimateur sans biais de faible variance.*

Exemple 4.1.2 *Soit le modèle paramétrique régulier* $(\mathcal{N}(\mu, \sigma^2), \theta = \mu \in \mathbb{R})$ *. Alors on calcule*

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

D'autre part (cf. Chapitre 2)

$$Var(\overline{X}_n) = \frac{\sigma^2}{n},$$

donc la moyenne empirique est un estimateur efficace pour μ*.*

Remarque 12 *La variance de* S_n^2 *est plus petite que la borne de Cramer-Rao :* $Var_\theta(S_n^2) = 2\sigma^4(n-1)/n^2 < 2\sigma^4/n = I_n(\sigma^2)^{-1}$*. Ce n'est pas en contradiction avec le théorème de Cramer-Rao car* S_n^2 *est biaisé ! L'estimateur* $S_n^{2\prime}$*, non biaisé, n'est lui pas efficace car de variance plus grande* $Var_\theta(S_n^{2\prime}) = 2\sigma^4/(n-1) > 2\sigma^4/n = I_n(\sigma^2)^{-1}$*. Pour comparer* S_n^2*, biaisé et de variance plus petite, et* $S_n^{2\prime}$ *non biaisé et de variance plus grande, il faut comparer leurs risques quadratiques. On trouve*

$$R(S_n^2, \sigma^2) = Var_\theta(S_n^2) + b_n(\sigma^2)^2 = \frac{2\sigma^4(n-1)}{n^2} + \left(\frac{\sigma^2}{n}\right)^2 = \frac{2\sigma^4}{n} - \frac{\sigma^4}{n^2}$$
$$< \frac{2\sigma^4}{n-1} = Var_\theta(S_n^{2\prime}) = R(S_n^{2\prime}, \sigma^2).$$

Dans le modèle gaussien, S_n^2 *est donc un meilleur estimateur que* $S_n^{2\prime}$*.*

4.2 Modèles de la famille exponentielle

Dans le cadre de l'estimation ponctuelle, l'objectif du statisticien est d'obtenir le meilleur estimateur possible du paramètre inconnu $\theta \in \Theta$. Le critère non asymptotique de la variance minimale garantit l'optimalité d'un estimateur sans biais parmi la classe des estimateurs sans biais. Il est possible de construire de tels estimateurs pour les modèles de la famille exponentielle.

4.2.1 Définitions et premières propriétés

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

Définition 4.2.1 *Un modèle $(P_\theta, \theta \in \Theta)$ est un modèle de la famille exponentielle s'il existe des fonctions à valeurs réelles $\theta \mapsto \alpha_j(\theta), \theta \mapsto c(\theta)$, $x \mapsto T_j(x)$ et $x \mapsto h(x)$ telles que la densité de P_θ soit de la forme*

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

Exemple 4.2.1 *Le modèle $(\mathcal{N}(\mu,\sigma^2), \theta = (\mu,\sigma^2) \in \mathbb{R}\times\mathbb{R}_+^*)$ est de la famille exponentielle :*

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

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

4.2.2 Notion d'identifiabilité

Dans le cadre de l'estimation statistique, la notion d'identifiabilité du modèle paramétrique est une condition naturelle, voir remarque ci-dessous. A l'étape de modélisation du problème, il faut autant que possible la respecter.

Définition 4.2.2 *Un modèle paramétrique $(P_\theta, \theta \in \Theta)$ est identifiable ssi l'application $\theta \mapsto P_\theta$ est injective.*

La notion d'identifiabilité dépend de la paramétrisation choisie :

Exemple 4.2.2 *Le modèle gaussien $(\mathcal{N}(0,\sigma^2), \theta = \sigma \in \mathbb{R}\setminus\{0\})$ n'est pas identifiable car à partir de la loi suivie par l'échantillon on ne distingue pas les cas $\theta = \sigma$ et $\theta = -\sigma$. Par contre pour la paramétrisation usuelle $\theta = |\sigma| > 0$ (l'écart type) il est bien identifiable (c.f. ci dessous pour une méthode effective pour prouver l'identifiabilité).*

Remarque 13 *L'hypothèse d'intifiabilité est équivalente à, pour tous* $\theta, \theta' \in \Theta$*,*

$$\nu(x \in \mathcal{X} : \ f(x,\theta) = f(x,\theta')) > 0 \Longrightarrow \theta = \theta'.$$

Supposons que $(P_\theta, \theta \in \Theta)$ *ne soit pas identifiable. Alors il existe* $\theta \neq \theta'$ *2 paramètres distincts tels que* $P_\theta = P_{\theta'}$*. Soit l'échantillon* $X_1, \ldots, X_n \sim P_\theta$ *où* θ *inconnu est le paramètre à estimer. Comme* $P_\theta = P_{\theta'}$*, l'information apportée par l'échantillon ne permet pas de distinguer* θ *de* θ'*.*

Pour un modèle paramétrique donné l'identifiabilité n'est pas facile à vérifier. Dans le cas de la famille exponentielle, il est possible de vérifier facilement qu'un modèle est identifiable avec le résultat suivant :

Proposition 4.2.1 *Si* (P_θ, Θ) *est un modèle de la famille exponentielle tel que la famille de fonctions* $(T_j(x))_{1 \leq j \leq r}$ *(définies sur le support* $\{x \in \mathcal{X} \ / \ f(x,\theta) > 0\}$*) soit affinement indépendante et tel que* $\alpha : \theta \to (\alpha_1(\theta), \ldots, \alpha_r(\theta))$ *soit injective alors ce modèle est identifiable.*

Remarque 14

- *Les famille de fonctions* (α_j) *et* (T_j) *ne sont pas déterminée de manière unique : on les identifie par rapport à l'expression de la densité. On choisit ces familles les plus simples possibles de manière à ce que le modèle soit identifiable.*
- *La famille de fonctions* $(f_1, ..., f_k)$ *est dite affinement indépendante ssi*

$$a_1 f_1 + \cdots + a_k f_k = a_{k+1} \implies a_1 = \cdots = a_k = a_{k+1} = 0.$$

- *Une famille réduite à une fonction* (f) *est affinement indépendante dès que* f *n'est pas constante sur leur domaine de définition.*
- *Une fonction* α *est injective si elle est continûment différentiable* (C^1) *et que sa matrice Jacobienne* $(\partial\alpha_i/\partial\theta_j)_{1 \leq i \leq r, 1 \leq j \leq d}$ *est continue et de rang* d $(r \geq d)$ *en tout point* $\theta \in \Theta$*. On dit alors que* α *est* C^1 *de Jacobienne de plein rang.*
- *Une fonction à valeur réelle* α *est injective si elle est continûment dérivable de dérivée non nulle.*

Proposition 4.2.2 *Soit* $X_1, \ldots, X_n \sim P_\theta$ *un échantillon issu d'un modèle de la famille exponentielle régulier vérifiant les hypothèses de la Propostion 4.2.1, alors*

$$T_n = \Big(\sum_{i=1}^n T_1(X_i), \ldots, \sum_{i=1}^n T_r(X_i) \Big)$$

est une statistique exhaustive appelée la <u>statistique exhaustive complète</u>*. Elle est unique à un facteur multiplicatif près.*

Démonstration : La densité de l'échantillon est de la forme :

$$f((x_1,\dots,x_n),\theta) = c(\theta)^n \prod_{i=1}^{n} h(x_i) \exp\left(\sum_{j=1}^{r} \alpha_j(\theta) \sum_{i=1}^{n} T_j(x_i)\right).$$

D'après le théorème de factorisation, on trouve donc la statistique exhaustive T_n pour le paramètre θ. Elle est unique à un facteur multiplicatif près car sinon on est en contradiction avec l'hypothèse d'indépendance affine. □

Tous les modèles classiques munis de leur paramétrisation classique sont identifiables :

Exemple 4.2.3

- *Dans le modèle gaussien* $(\mathcal{N}(\mu,\sigma^2), \theta = (\mu,\sigma^2) \in \mathbb{R}\times\mathbb{R}_+^*)$*, on est dans la famille exponentielle avec*

$$\alpha_1(\theta) = \frac{\mu}{\sigma^2}, \qquad \alpha_2(\theta) = -\frac{1}{2\sigma^2}, \qquad T_1(x) = x, \quad \text{et} \quad T_2(x) = x^2.$$

  On vérifie que (T_1,T_2) *est une famille de fonctions affinements indépendantes (en choisissant par exemple* $x = 0,1$ *et* -1*). De plus* α *est* C^1 *car de Jacobienne*

$$J(\alpha(\theta)) = \begin{pmatrix} \frac{1}{\sigma^2} & -\frac{\mu}{\sigma^4} \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix},$$

  matrice de déterminant $\sigma^{-6}/2$ *non nulle donc de rang 2 donc de plein rang. Le modèle* (P_θ,Θ) *est donc identifiable et la statistique exhaustive complète vaut*

$$T_n = \Big(\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} X_i^2\Big).$$

- *Dans le modèle gaussien* $(\mathcal{N}(\mu,\sigma^2), \theta = \sigma^2 > 0)$ *(*μ *est connu) on a la densité*

$$f(x,\theta) = \frac{1}{\sqrt{2\pi\theta}} \exp\Big(-\frac{(x-\mu)^2}{2\theta}\Big).$$

  On reconnaît un modèle de la famille exponentiel avec $\alpha_1(\theta) = -1/(2\theta)$ *et* $T_1(x) = (x-\mu)^2$ *(valable car* μ *est connu). Ces 2 fonctions sont non constantes et* α *est différentiable injective donc le modèle est bien identifiable. La statistique exhaustive complète du modèle est* $T_n = \sum_{i=1}^{n}(X_i-\mu)^2$*.*
- *Soit* $(\mathcal{B}(m,p), 0 < \theta = p < 1)$*, on a alors*

$$f(x,\theta) = C_m^x (1-p)^m \exp\Big(x \log\Big(\frac{p}{1-p}\Big)\Big).$$

On reconnaît un modèle de la famille exponentielle avec $c(\theta) = (1-\theta)^m$, $h(x) = C_m^x$, $T_1(x) = x$ *et* $\alpha_1(\theta) = \log(p/(1-p))$. *La fonction* T_1 *n'est pas constante et la fonction* α_1 *est dérivable de dérivée continue* $1/(p(1-p)) \neq 0$ *donc de rang 1 donc de plein rang. Le modèle* (P_θ, Θ) *est identifiable et sa statistique exhaustive complète est* $\sum_{i=1}^n X_i$.

– *Soit* $(\gamma(\alpha,\beta), \theta = (\alpha,\beta) \in]0,\infty[^2)$, *alors pour* $x > 0$, *on a :*

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

On est bien dans la famille exponentielle où on identifie $\alpha(\theta) = (-\beta, \alpha-1)$ *et* $(T_1,T_2)(x) = (x, \log(x))$. *Comme* $J_\theta(\alpha(\theta)) = Id_2$ *continue de plein rang le modèle* (P_θ,Θ) *est identifiable et sa statistique exhaustive complète est*

$$T_n = \Big(\sum_{i=1}^n X_i, \sum_{i=1}^n \log(X_i)\Big).$$

4.3 Estimation non asymptotique dans la famille exponentielle

4.3.1 Théorème de Lehmann-Scheffé

Il est possible de déterminer un estimateur sans biais de variance minimale dans un modèle de la famille exponentielle identifiable. Soit (P_θ,Θ) un modèle de la famille exponentielle identifiable. Rappelons que

$$S_n = \left(\sum_{i=1}^n T_1(X_i), \ldots, \sum_{i=1}^n T_r(X_i)\right)$$

est une statistique exhaustive. Afin de pouvoir parler de variance, nous allons supposer le modèle régulier :

Proposition 4.3.1 *Un modèle de la famille exponentielle vérifiant les hypothèses de la Proposition 4.2.1 et tel que* α *est 2 fois continûment différentiable et* $\mathbb{E}_\theta(T_j^2(X)) < \infty$ *pour tout* $1 \leq j \leq r$ *alors le modèle* (P_θ,Θ) *est identifiable et régulier.*

Le principal résultat de ce chapitre est le suivant

Théorème 4.3.1 (Théorème de Lehmann-Scheffé) *Soit un modèle de la famille exponentielle identifiable et régulier vérifiant les hypothèses de la Proposition 4.3.1. L'* unique *estimateur de* θ *sans biais de variance minimale est l'unique fonction de la statistique exhaustive complète* T_n *sans biais.*

Exemple 4.3.1

- *Dans le modèle gaussien* $(\mathcal{N}(\mu,\sigma^2), \theta=(\mu,\sigma^2)\in\mathbb{R}\times\mathbb{R}_+^*)$ *on déduit du Théorème de Lehmann-Scheffé que* $(\overline{X}_n, S_n^{2\prime})$*, fonction de* $T_n=(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$*, est l'estimateur sans biais de variance minimale.*
- *Dans le modèle gaussien* $(\mathcal{N}(\mu,\sigma^2), \theta=\sigma^2>0)$ *(μ est connu), l'estimateur* $\overline{(X-\mu)^2}_n$ *est l'unique estimateur sans biais de variance minimale car fonction de* $T_n=\sum_{i=1}^n (X_i-\mu)^2$*.*
- *Soit* $(\mathcal{B}(m,p), 0<\theta=p<1)$ *identifiable avec* $\sum_{i=1}^n X_i$ *la statistique exhaustive complète. Donc* $\overline{X}_n$ *est l'estimateur de variance minimale.*

4.4 Efficacité et modèles de la famille exponentielle

Dans un modèle de la famille exponentiel identifiable et régulier l'information de Fisher est bien définie ainsi que la borne de Cramer-Rao. Il est donc naturel de comparer la variance d'un estimateur avec cette borne. Si elle sont égales, l'estimateur est efficace et c'est aussi l'unique estimateur sans biais de variance minimale. Si ce n'est pas le cas, l'estimateur peut tout de même être de variance minimale ; le modèle n'admet alors pas d'estimateur efficace.

Exemple 4.4.1 *Dans le cas gaussien* $(\mathcal{N}(\mu,\sigma^2), \theta=\sigma^2>0)$ *:*

- *Dans le modèle gaussien* $(\mathcal{N}(\mu,\sigma^2), \theta=(\mu,\sigma^2)\in\mathbb{R}\times\mathbb{R}_+^*)$ *identifiable* $(\overline{X}_n, S_n^{2\prime})$ *est l'estimateur sans biais de variance minimale. Il n'est pas efficace car*

$$Var_\theta(S_n^{2\prime}) = \frac{2\sigma^2}{n-1} > \frac{2\sigma^2}{n} = (I_n^{-1}(\theta))_{2,2}.$$

- *Dans le modèle gaussien* $(\mathcal{N}(\mu,\sigma^2), \theta=\sigma^2>0)$ *(μ est connu) identifiable* $\overline{(X-\mu)^2}_n$ *est l'estimateur de variance minimale. Il est efficace car* $Var_\theta(\overline{(X-\mu)^2}_n) = (\mu_4-\theta^2)/n = 2\theta^2/n$*.*
- *Soit* $(\mathcal{B}(m,p), 0<\theta=p<1)$ *identifiable avec* $\overline{X}_n$ *l'estimateur de variance minimale de variance* $m\theta(1-\theta)/n = I_n^{-1}(\theta)$ *donc efficace.*

La notion d'efficacité est souvent trop forte et n'est utile que dans un petit nombre de modèles. On lui préfère celle de variance minimale dans le cas d'un modèle de la famille exponentielle ou celle d'efficacité asymptotique dans le cas d'un modèle régulier, c.f. chapitre suivant.

Chapitre 5

Approche asymptotique

5.1 Critères asymptotiques

Nous allons voir que l'asymptotique simplifie souvent la comparaison de divers estimateurs. En particulier ce cadre permet de s'affranchir du cadre d'estimateur sans biais.

5.1.1 Estimateur asymptotiquement sans biais

Définition 5.1.1 *Un estimateur T_n de θ est dit asymptotiquement sans biais si*

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

De nombreux estimateurs biaisé sont asymptotiquement sans biais. Cette hypothèse est souvent plus réaliste dans des cas pratiques.

Exemple 5.1.1 *Supposons que X soit de carré intégrable, i.e. $Var\,(X) = \Sigma^2 < \infty$. La variance empirique*

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

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

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

Dans le cadre asymptotique, différents modes de convergence de l'e.a. T_n vers θ sont envisageables. Le biais n'est pas un mode de convergence classique en probabilité, on lui préfère les convergences en probabilité ou presque sûrement.

5.1.2 Estimateur convergent

Définition 5.1.2 *Un estimateur T_n est convergent (ou consistant) s'il converge en probabilité vers θ*

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

On notera $T_n \xrightarrow{P} \theta$ (en omettant l'indice θ pour la loi P).

Cette notion est souvent plus forte que la notion d'asymptotiquement sans biais :

Proposition 5.1.1 *Un estimateur T_n asymptotiquement sans biais qui vérifie en plus $Tr(Var_\theta(T_n)) \to 0$ est convergent en moyenne quadratique (dans L^2), i.e. son risque quadratique $R(T_n, \theta)$ tend vers 0.*
Réciproquement, un estimateur T_n convergent et tel qu'il existe X intégrable vérifiant $\|T_n\| \leq X$ est asymptotiquement sans biais.

Démonstration : Pour le premier point, d'après la décomposition biais variance, l'estimateur T_n est donc convergent par comparaison des modes de convergence.
Pour le second point, on utilise le théorème de convergence dominé.

Définition 5.1.3 *Un estimateur T_n est fortement convergent (ou consistant) s'il converge presque sûrement (p.s.) vers θ*

$$P_\theta(\lim_{n\to\infty} T_n = \theta) = 1.$$

On notera $T_n \xrightarrow{p.s.} \theta$.

Un estimateur fortement convergent est convergent d'après la comparaison des différents modes de convergence.

5.1.3 Efficacité asymptotique d'un estimateur

Cette notion n'est valable que pour les estimateurs asymptotiquement sans biais (donc pour la plupart des estimateurs convergents et, a fortiori, fortement convergents).

Lorsqu'on compare deux estimateurs convergents dans un cadre asymptotique, il est naturel de comparer les variances de leurs lois asymptotiques respectives, qui est en générale la loi normale :

Définition 5.1.4 *Un estimateur T_n de θ est asymptotiquement normal si il satisfait un TLC : il existe $\Sigma^2(\theta)$ une matrice symétrique positive de dimension $d \times d$ telle que*

$$\sqrt{n}(T_n - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}_d(0_d, \Sigma^2(\theta)).$$

La matrice de variance-covariance Σ ne dépend pas de n. On l'appelle abusivement la variance asymptotique de T_n.

Proposition 5.1.2 *Un estimateur asymptotiquement normal est nécessairement fortement convergent.*

Démonstration : Soit T_n un estimateur asymptotiquement normal de variance asymptotique définie positive, i.e. tel que

$$\sqrt{n}\Sigma(\theta)^{-1}(T_n - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}_d(0_d, I_d).$$

Or, la convergence en loi étant stable par transformation continue, si on applique la fonction $x \to \|x\|^2 = x^T x$ continue on obtient, comme $Z = \|N\|^2 \sim \chi^2_d$ si $N \sim \mathcal{N}_d(0_d, I_d)$ par définition :

$$n(T_n - \theta)^T \Sigma^2(\theta)^{-1}(T_n - \theta) \xrightarrow{\mathcal{L}} \chi^2_d.$$

La convergence en loi implique pour tout $\epsilon > 0$ l'équivalence pour n grand

$$\mathbb{P}(\sqrt{n}(T_n - \theta)^T \Sigma^2(\theta)^{-1}(T_n - \theta) \geq \epsilon) \approx \mathbb{P}(\|N\|^2 \geq \sqrt{n}\epsilon).$$

Mais d'après la densité d'une χ^2_d on a pour n grand

$$\mathbb{P}(N^2 \geq \sqrt{n}\epsilon) \leq (\sqrt{n}\epsilon)^{d/2-1} \exp(-\sqrt{n}\epsilon d/2)$$

qui est elle même une série convergente. Par le théorème de convergence dominée, on trouve donc pour tout $\epsilon > 0$

$$\sum_{n \geq 0} \mathbb{P}(\sqrt{n}(T_n - \theta)^T \Sigma^2(\theta)^{-1}(T_n - \theta) \geq \epsilon) < +\infty$$

et on conclut par Borel-Cantelli que $\sqrt{n}(T_n - \theta)^T \Sigma^2(\theta)^{-1}(T_n - \theta) \xrightarrow{p.s.} 0$. Enfin, il est facile de voir que pour tout θ on a $N_\theta(u) = u^T \Sigma^2(\theta)^{-1} u$ qui est une norme vectorielle sur $\mathbb{R}^d$. Comme toutes les normes sont équivalentes, la convergence p.s. vers 0 du vecteur $T_n - \theta$ pour cette norme implique sa convergence p.s. vers 0 pour la norme usuelle, autrement dit T_n est bien fortement convergent.

Définition 5.1.5 *Soient T_n et T'_n 2 estimateurs asymptotiquement normaux de θ. Alors T_n est asymptotiquement plus efficace que T'_n si, notant Σ et Σ' leurs variances asymptotiques respectives, on a*

$$\Sigma(\theta) \leq \Sigma'(\theta), \quad \forall\, \theta \in \Theta \quad \text{et} \quad \exists \theta' \in \Theta, \quad \Sigma(\theta') < \Sigma'(\theta').$$

Définition 5.1.6 *Un estimateur est asymptotiquement efficace lorsqu'il est asymptotiquement normal et que sa matrice de variance covariance limite* $\Sigma(\theta) = I^{-1}(\theta)$*, i.e. il atteint la borne de Cramer-Rao asymptotique.*

Exemple 5.1.2 *Dans le modèle paramétrique régulier* $(\mathcal{N}(\mu, \sigma^2), \theta = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+^*)$*, la variance empirique* S_n^2 *est un estimateur asymptotiquement efficace de* σ^2*. En effet, l'information de Fisher pour* σ^2*, notée* $I_n(\sigma^2)$ *vaut* $I_n(\theta)^{(2,2)}$*, soit* $nI(\theta)^{(2,2)} = n/(2\sigma^4)$*. D'autre part, on a vu que dans le cas gaussien* $Var_\theta(S_n^2) = 2\sigma^4(n-1)/n^2$*. D'où le résultat pour* S_n^2*. Il en va de même pour* $S_n^{2'}$ *car* $Var_\theta(S_n^{2'}) = n^2/(n-1)^2\, Var_\theta(S_n^2)$*. Il n'est pas possible de distinguer* S_n^2 *et* $S_n^{2'}$ *selon un critère asymptotique : ils sont tous les 2 aussi bons, à savoir asymptotiquement efficaces.*

Remarque 15

- *La convergence en loi n'entraîne pas nécessairement la convergence des matrices de variance-covariance donc un estimateur peut être asymptotiquement efficace sans pour autant avoir*

$$\lim_{n\to\infty} n\, Var_\theta(T_n) I(\theta) = Var_\theta(T_n) I_n(\theta) = I_d.$$

  En particulier il existe des estimateurs dont la matrice de variance-covariance asymptotique est plus petite que la borne de Cramer-Rao asymptotique.
- *Un estimateur efficace pour* nn_0 *avec* n_0 *fixé est asymptotiquement efficace.*

5.2 Les Z-estimateurs

Les Z-estimateurs sont des généralisations des moments empiriques. On donne la définition formelle puis on étudie des cas particuliers.

Définition 5.2.1 *Soit une fonction*

$$\Phi : \mathcal{X} \times \Theta \to \mathbb{R}^d$$

intégrable par rapport à P_θ *pour tout* $\theta \in \Theta$ *et telle que*

$$\mathbb{E}_\theta(\Phi(X, \theta)) = 0_d.$$

Tout estimateur $T_n = T(X_1, \ldots, X_n)$ *qui vérifie*

$$\frac{1}{n}\sum_{i=1}^{n} \Phi(X_i, T_n) = 0_d$$

est appelé un Z-estimateur.

5.2.1 Les moments empiriques

Le paramètre inconnu θ est un moment lorsque $\theta = \mathbb{E}_\theta(X^r)$ dans le cas $\mathcal{X} = \mathbb{R}$. Les moments empiriques d'ordre r sont des Z-estimateurs pour la fonction $\Phi(x, \theta) = x^r - \theta$.

5.2.2 La méthode des moments

Supposons qu'il existe une fonction $g : \Theta \to \mathbb{R}^d$ inversible et d-moments (non centrés) m_{i_j}, $1 \leq j \leq d$ tels que

$$g(\theta) = (m_{i_1}, \ldots, m_{i_d})^T.$$

L'estimateur obtenu par la méthode des moments (MM) est alors donné par la formule

$$T_n = g^{-1}(M_n^{i_1}, \ldots, M_n^{i_d}).$$

C'est un Z-estimateur car solution du système

$$\frac{1}{n}\sum_{i=1}^{n} \Phi(X_i, T_n) = 0_d$$

avec $\Phi(x, \theta) = (x^{i_1}, \ldots, x^{i_d})^T - g(\theta)$.

Exemple 5.2.1 *Soit le modèle exponentiel $(\mathcal{E}(\theta),]0, \infty[)$. On sait que $\mathbb{E}_\theta(X) = \theta^{-1}$ et que $\mathbb{E}_\theta(X^2) = 2\theta^{-2}$. La méthode des moments fournit donc 2 estimateurs distincts T_n^1 et T_n^2 de θ, selon qu'on utilise le moment d'ordre 1 avec $g_1(x) = x^{-1}$ ou le moment d'ordre 2 avec $g_2(x) = 2x^{-2}$. On obtient*

$$T_n^1 = g_1^{-1}(M_n^1) = \frac{1}{\overline{X}_n} \quad et \quad T_n^2 = g_2^{-1}(M_n^2) = \frac{2}{\sqrt{\overline{X^2}_n}}.$$

Ce sont 2 estimateurs biaisés, pour les comparer il faut comparer leurs risques quadratiques respectifs.

La méthode des moments permet très facilement de construire des estimateurs pour des lois qui ont des bonnes propriétés de moments. Elle ne peut pas être utilisée si $X \sim P_\theta$ n'est pas intégrable :

Exemple 5.2.2 *Soit le modèle de Cauchy $(P_\theta, \mathbb{R})$ tel que*

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

Alors $xf(x, \theta) \approx \theta(x - \theta)^{-2}$ au voisinage de $\theta \neq 0$, ou $xf(x, \theta) \approx 1/x$ si $\theta = 0$ n'est pas intégrable sur $\mathbb{R}$. On en déduit que X n'est pas intégrable, m_r pour $r \geq 1$ n'existent pas et la méthode des moments est inutilisable.

5.2.3 La méthode des moments généralisés

Tout Z-estimateur qui n'est pas obtenu par la méthode des moments est un estimateur obtenu par la méthode dite méthode des moments généralisés (MMG). Reprenons l'exemple du modèle de Cauchy

Exemple 5.2.3 *Soit le modèle de Cauchy* $(P_\theta, \mathbb{R})$ *tel que*

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

On remarque que, ou $signe(x) = 1$ *ssi* $x > 0$, $signe(x) = -1$ *sinon, on a*

$$\mathbb{E}_\theta(signe(X)) = \int_{\mathbb{R}} signe(x) f(x,\theta)dx = \frac{1}{\pi}\left[\int_{-\theta}^{\infty} \frac{du}{1+u^2} - \int_{-\infty}^{-\theta} \frac{du}{1+u^2}\right]$$

en posant les changement de variables $u = x-\theta$. *Comme la primitive de* $(1+u^2)^{-1}$ *est* $Arctan(u)$, *on trouve*

$$\mathbb{E}_\theta(signe(X)) = \frac{1}{\pi}\left[\frac{\pi}{2} - \arctan(-\theta) - \arctan(-\theta) - \frac{\pi}{2}\right] = 2\arctan(\theta)/\pi.$$

En posant $\Phi(x,\theta) = signe(x) - 2\arctan(\theta)/\pi$ *on trouve le* Z*-estimateur de* θ *:*

$$T_n = \tan\Big(\frac{\pi}{2n}\sum_{i=1}^{n} signe(X_i)\Big).$$

5.2.4 Extension : les quantiles empiriques

On rappelle la définition d'un quantile :

Définition 5.2.2 *Le quantile d'ordre* $\alpha \in]0,1[$ *de* X *est noté* q_α *et est donné par la formule*

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

Remarque 16 *Dans le cas discret* $X \in \{x_i\}_{i\in I} = \mathcal{X}$ *alors* $q_\alpha = \inf\{x_i,\ i \in I \mid F(x_i) \geq \alpha\}$. *En particulier* $q_\alpha \in \mathcal{X}$.

Soit un modèle paramétrique tel que le paramètre d'intérêt inconnu soit $\theta = q_\alpha$. On a q_α qui est le plus petit réel tel que $F_\theta(q_\alpha) \geq \alpha$ qu'on réécrit :

$$\mathbb{E}_\theta(\,1_{X\leq\theta}) \geq \alpha.$$

En posant $\Phi(x,\theta) = 1_{x\leq\theta} - \alpha$, on obtient que θ est l'infimum des points a qui vérifient

$$\mathbb{E}_\theta \Phi(X,a) \geq 0.$$

Contrairement au cas des Z-estimateurs, l'égalité n'a pas forcément lieu.

Exemple 5.2.4 *Soit le modèle $P = \mathcal{B}(p)$ avec $p \in]0,1[$ inconnu, on s'intéresse à $\theta = q_{0,5}$ la médiane. Si $1 > p > 1/2$ alors par définition $\theta = 1$ et $\mathbb{E}_\theta \Phi(X, \theta) = 1 \neq 0,5$.*

Par extension, on estime q_α par T_n qui réalise l'infimum des a

$$\frac{1}{n}\sum_{i=1}^{n} 1_{X_i \leq a} \geq \alpha.$$

On appelle T_n le quantile empirique :

Définition 5.2.3 *Le quantile empirique vaut $T_n = X_{(\lceil n\alpha \rceil)}$ où $(X_{(1)}, \ldots, X_{(n)})$ est l'échantillon ordonné et $\lceil y \rceil$ est le plus petit entier plus grand que y.*

Le quantile empirique est une statistique d'ordre.

Dans le cas où F est absolument continue, le quantile empirique T_n est un estimateur fortement convergent et asymptotiquement normal :

Théorème 5.2.1 *Soient $(P_\theta, \theta = q_\alpha \in \mathbb{R})$ un modèle paramétrique tel que P_θ soit absolument continue, i.e. admette une densité notée f_θ par rapport à la mesure de Lebesgue. Soit $\alpha \in]0,1[$ et soit T_n le quantile empirique d'ordre α, alors il est asymptotiquement normal*

$$\sqrt{n}(T_n - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \frac{\alpha(1-\alpha)}{f_\theta^2(\theta)}\right).$$

5.3 Les M-estimateurs

On commence par donner la définition formelle des M-estimateurs puis on étudie des cas particuliers.

Définition 5.3.1 *Soit une fonction*

$$\Psi : \mathcal{X} \times \Theta \to \mathbb{R}$$

intégrable par rapport à P_θ pour tout $\theta \in \Theta$ et telle qu'on ait

$$\arg\max_{a \in \Theta} \ \mathbb{E}_\theta(\Psi(X, a)) = \theta.$$

Tout estimateur $T_n = T(X_1, \ldots, X_n)$ qui vérifie

$$\frac{1}{n}\sum_{i=1}^{n} \Psi(X_i, T_n) = \max_{a \in \Theta} \frac{1}{n}\sum_{i=1}^{n} \Psi(X_i, a)$$

est un M-estimateur.

5.3.1 Paramètre de localisation

Soit $(P_\theta, \theta \in \Theta = \mathbb{R})$ telle que θ soit le paramètre de localisation : il existe une fonction de répartition F telle que
- si $X \sim P_\theta$ alors $P_\theta(X \leq x) = F(x - \theta)$ pour tout $x \in \mathcal{X} = \mathbb{R}$ et tout $\theta \in \mathbb{R}$,
- si $X \sim F$ alors $\mathbb{E}(X) = 0$ et $\mathbb{E}(X^2) < \infty$.

Exemple 5.3.1 *Dans le cas du modèle Gaussien $\mathcal{N}(\mu, \sigma^2)$, μ est le paramètre de localisation quelque soit la valeur de σ^2.*

On a $a \mapsto \mathbb{E}_\theta(X - a)^2$ qui est bien définie et de dérivée première $-2\mathbb{E}_\theta(X) + 2a$, de dérivée seconde 2. C'est une fonction convexe qui a un unique minimum en $a = \mathbb{E}_\theta(X) = \int_{\mathbb{R}} x dP_\theta(x) = \int_{\mathbb{R}}(x + \theta)dP(x) = \theta$. En notant $\Psi(x, \theta) = -(x - \theta)^2$ on obtient ainsi un M-estimateur T_n de θ vérifiant

$$\sum_{i=1}^{n}(T_n - X_i)^2 = \max_{a\in\mathbb{R}} \sum_{i=1}^{n}(a - X_i)^2.$$

Remarque 17 *On obtient $T_n = \overline{X}_n$ comme estimateur de $\mathbb{E}_\theta(X) = \theta$. Attention, dans le cas du modèle exponentiel $\mathcal{E}(\lambda)$ alors $\mathbb{E}_\lambda(X) = 1/\lambda$ n'est pas un paramètre de localisation car il n'existe pas de fonction F telle que $P_\lambda(X \leq x) = F(x - 1/\lambda)$.*

5.3.2 Estimateur des moindres carrés

On se place dans le cadre d'un modèle appelé modèle linéaire (simple) où $\mathcal{X} = \mathbb{R}^2$, et on note les couples d'observations $(X_i, Y_i)_{1\leq i\leq n}$ issus du couple (X, Y) de carré intégrable et qui satisfait la relation

$$Y = b_1 + b_2 X + \varepsilon,$$

avec ε une v.a. centrée de variance σ^2 et indépendant de X. Les couples (X_i, Y_i) sont indépendants entre eux mais les Y_i dépendent de X_i ! Le paramètre d'intérêt est $\theta = (b_1, b_2) \in \Theta = \mathbb{R}^2$ qu'on estime avec $T_n = (T_n^{(1)}, T_n^{(2)})^T$ obtenu par la méthode des moindres carrés :

$$\sum_{i=1}^{n}(Y_i - T_n^{(1)} - T_n^{(2)} X_i)^2 = \min_{(a_1,a_2)\in\mathbb{R}^2} \sum_{i=1}^{n}(Y_i - a_1 - a_2 X_i)^2.$$

T_n est un M-estimateur associé à la fonction $\Psi((x, y), (b_1, b_2)) = -(y - b_1 - b_2 x)^2$ en vérifiant bien que le critère $(a_1, a_2) \mapsto \mathbb{E}_\theta(Y - a_1 - a_2 X)^2$ est minimal pour $(a_1, a_2) = \theta = (b_1, b_2)$.

5.3.3 Maximum de vraisemblance

C'est le plus important des M-estimateurs car il est associé à un choix de Ψ en accord avec la théorie de l'information, i.e. qui assure qu'en choisissant le maximum pour $\Psi = n^{-1}\sum_{i=1}^{n}\Psi(a, X_i)$ on garde le maximum d'information (voir les propriétés du maximum de vraisemblance dans le chapitre suivant).

Définition 5.3.2 *On appelle vraisemblance de l'échantillon $X_1, \ldots, X_n \sim P_\theta$ en $a \in \Theta$ la v.a. à valeurs dans $\mathbb{R}^+$ définie par*

$$L_n(a) = f((X_1, \ldots, X_n), a),$$

i.e. la densité $f((x_1, \ldots, x_n), a)$ exprimée en les observations $X_i \sim P_\theta$.

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

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

Définition 5.3.3 *Soit $L_n(a)$ la vraisemblance au point $a \in \Theta$. On appelle estimateur du maximum de vraisemblance (EMV) la statistique $T_n = T(X_1, \ldots, X_n)$ telle que*

$$L_n(T_n) = \max_{a\in\Theta} L_n(a).$$

Sous cette forme générale, l'EMV n'est pas un M-estimateur dans le sens où le critère à maximiser s'écrit sous forme d'un produit et non d'une somme. On déduit des propriétés de l'EMV de sa définition :

Propriété 3

1. *L'EMV n'existe pas toujours.*
2. *Il n'y aucune raison pour que l'EMV soit sans biais.*
3. *L'EMV n'a aucune raison d'être unique.*

Exemple 5.3.2 *Soit $(\mathcal{U}[0, \theta], \theta > 0)$ alors*

$$L_n(\theta) = \prod_{j=1}^{n} \frac{1}{\theta} 1_{\ [0,\theta]}(X_j) = \frac{1}{\theta^n} 1_{\ 0\le X_{(1)}\le X_{(n)}\le\theta} = \frac{1}{\theta^n} 1_{[\ \sup_{1\le j\le n} X_j, \infty[}(\theta)$$

et donc on trouve l'EMV $T_n = \sup_{1\le j\le n} X_j$ directement car $1_{0\le \inf_{1\le j\le n} X_j} = 1$ p.s.. On peut montrer que $T_n/\theta \sim Beta(n, 1)$, i.e. la loi de la variable aléatoire T_n admet pour densité

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

Il s'en suit que $\mathbb{E}_\theta(T_n/\theta) = n/(n+1)$ et $b_\theta(T_n) = E(T_n) - \theta = -\theta/(n+1) \neq 0$ donc l'EMV est ici biaisé.

Exemple 5.3.3 *Soient* $(\mathcal{U}[\theta, \theta+1], \theta > 0)$ *alors tout estimateur* T_n *compris entre* $\sup_{1\leq i\leq n} X_i - 1$ *et* $\inf_{1\leq i\leq n} X_i$ *est un EMV de* θ .

Proposition 5.3.1 *Si le modèle* (P_θ, Θ) *vérifie l'hypothèse (H1) alors l'EMV est un M-estimateur avec* $\Psi(x, a) = \log f(x, a)$.

Démonstration : L'hypothèse (H1) étant satisfaite, le support $S = \{x \in \mathcal{X} \,/\, f(x, \theta) > 0\}$ ne dépend pas de θ. Par définition du support, les observations $X_i \in S$ car $X_i \sim P_\theta$ et donc $f(X_i, a) > 0$ pour tout $1 \leq i \leq n$. Pour tout $a \in \Theta$ la vraisemblance $L_n(a) = \prod_{i=1}^n f(X_i, a)$ est donc strictement positive. On peut donc passer au logarithme, le logarithme étant croissante, l'EMV est aussi le maximum de $\Psi(x, \theta)$. Reste à vérifier que $\mathbb{E}_\theta(\Psi(x, a))$ réalise un maximum global en θ. Soit $a \in \Theta$, par définition

$$\mathbb{E}_\theta(\Psi(x, a)) = \mathbb{E}_\theta(\log f(X, a)) = \int_{\mathcal{X}} \log f(x, a) f(x, \theta) d\nu(x),$$

d'où

$$\mathbb{E}_\theta(\Psi(X, \theta)) - \mathbb{E}_\theta(\Psi(X, a)) = \int_{\mathcal{X}} \log\left(\frac{f(x, \theta)}{f(x, a)}\right) f(x, \theta) d\nu(x).$$

La fonction $x \mapsto -\log(x)$ étant convexe, on utilise l'inégalité de Jensen et on trouve

$$\mathbb{E}_\theta(\Psi(X, \theta)) - \mathbb{E}_\theta(\Psi(X, a)) \geq -\log\left(\int_{\mathcal{X}} \frac{f(x, a)}{f(x, \theta)} f(x, \theta) d\nu(x)\right) = -\log(1) = 0$$

donc θ est un maximum global. □

L'EMV est un M-estimateur qui maximise la fonction qui à a associe

$$\frac{1}{n} \sum_{i=1} \log f(X_i, a).$$

Définition 5.3.4 *On appelle fonction de log-vraisemblance la fonction* l_n *qui à* $a \in \Theta$ *associe*

$$l_n(a) = -\frac{1}{n} \sum_{i=1} \log f(X_i, a).$$

L'EMV est le minimisateur de la fonction de log-vraisemblance, issue du critère $a \mapsto -\mathbb{E}_\theta(\log f(X, a))$. Par définition, si le modèle est régulier, l'espérance de la matrice Hessienne de la fonction de log-vraisemblance au point θ est l'information de Fisher :

$$\mathbb{E}_\theta[\mathbb{H}_\theta(l_n(\theta))] = I(\theta) > 0.$$

5.4 Comparaison des Z et M-estimateurs

Sous des hypothèse de dérivabilité de la fonction Ψ, on remarque que tout M-estimateur est un Z-estimateur associé à la fonction $\Phi(x,\theta) = \nabla_\theta \Psi(x,\theta)$. Nous nous bornons donc à l'étude asymptotique des Z-estimateurs.

Les propriétés asymptotiques des Z-estimateurs (convergence, normalité asymptotique) sont donnés sous des les hypothèses loi limite des Z-estimateurs. On appelle $\Phi : \mathcal{X} \times \Theta \to \mathbb{R}^d$ une fonction vérifiant

$$\mathbb{E}_\theta(\Phi(X,\theta)) = 0_d \quad \text{pour tout} \quad \theta \in \Theta.$$

(i) Pour tout $\theta \in \Theta$, on a $\mathbb{E}_\theta \sup_{a\in\Theta} \|\Phi(a,X)\| < \infty$,

(ii) Pour tout $\epsilon > 0$, $\inf_{\|a-\theta\|>\epsilon} \|\mathbb{E}_\theta(\Phi(X,a))\| > 0$

(iii) Pour tout $a \in \Theta$ il existe un ouvert $V_a \subseteq \Theta$ contenant a et une fonction borélienne $g(x)$ tels que, pour tout $a \in V_a$

$$\|J_a\Phi(x,a)\| \le g(x), \qquad \|d_a(J_a\Phi(x,a))\| \le g(x) \qquad \text{et} \qquad \mathbb{E}_\theta(g(X)) < \infty,$$

où $d_a(J_a\Phi(x,a))$ est la différentielle de la matrice Jacobienne de $a \mapsto \Phi(x,a) \in \mathbb{R}^d$.

(iv) Pour tout $\theta \in \Theta$, on a $\mathbb{E}_\theta(\|\Phi(X,\theta)\|^2) < \infty$.

Théorème 5.4.1 *Sous les conditions de loi limite des Z-estimateurs, le Z-estimateur T_n solution de $\sum_{i=1}^n \Phi(X_i,T_n) = 0_d$ est asymptotiquement normal*

$$\sqrt{n}(T_n - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}_d(0_d, V_\Phi(\theta))$$

avec

$$V_\Phi(\theta) = (\mathbb{E}_\theta[J_\theta\Phi(X,\theta)])^{-1}\, Var_\theta[\Phi(X,\theta)](\mathbb{E}_\theta[J_\theta\Phi(X,\theta)])^{-1^T}.$$

Remarque 18

1. *Si la fonction $\theta \mapsto \nabla_\theta\Psi(x,\theta)$ vérifie les conditions de loi limite des Z-estimateurs, alors le M-estimateur correspondant est asymptotiquement normal*

$$\sqrt{n}(T_n - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}_d(0_d, V_\Psi(\theta))$$

avec

$$V_\Psi(\theta) = (\mathbb{E}_\theta[\mathbb{H}_\theta\Psi(X,\theta)])^{-1}\, Var_\theta[\nabla_\theta\Psi(X,\theta)](\mathbb{E}_\theta[\mathbb{H}_\theta\Psi(X,\theta)])^{-1^T}.$$

2. *Par abus de notation, on note parfois $T_n = \hat{\theta}_n^{MM}$ l'estimateur et l'estimation obtenus par la MM, $T_n = \hat{\theta}_n^{GMM}$ l'estimateur et l'estimation obtenus par la MM généralisés et $T_n = \hat{\theta}_n^{MV}$ l'estimateur et l'estimation obtenus par le MV.*

Dans de nombreux exemples, il est préférable d'utiliser les résultats connus sur les statistiques empiriques (c.f. chapitre 2) qui interviennent dans les Z-estimateurs de type moments ou moments généralisés :

Exemple 5.4.1 *Soit le modèle Gamma* $(\gamma(p,\lambda), \theta = (p,\lambda) \in]0,\infty[^2)$. *On sait que*

$$\mathbb{E}_\theta(X) = m_1 = \frac{p}{\lambda} \qquad et \qquad \mathbb{E}_\theta(X^2) = m_2 = \frac{p(p+1)}{\lambda^2}$$

et donc $\Phi(x,\theta) = (x - p/\lambda, x^2 - p(p+1)/\lambda^2)^T$. *On résout le système et on trouve* $p = m_1^2/(m_2 - m_1^2)$ *et* $\lambda = m_1/(m_2 - m_1^2)$ *d'où*

$$T_n = \left(\frac{(\overline{X}_n)^2}{S_n^2}, \frac{\overline{X}_n}{S_n^2} \right).$$

On utilise directement les résultats sur la moyenne et la variance empirique plutôt que de vérifier les conditions de loi limite des Z-estimateurs. On sait que

$$\sqrt{n}\left(\begin{pmatrix} \overline{X}_n \\ S_n^2 \end{pmatrix} - \begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix} \right) \xrightarrow{\mathcal{L}} \mathcal{N}_2\left(0_2, \begin{pmatrix} \sigma^2 & \mu_3 \\ \mu_3 & \mu_4 - \sigma^4 \end{pmatrix} \right).$$

Reste à calculer μ_3 *et* μ_4 *dans le cas Gamma. On a facilement les moments (non centrés)* $m_3 = \mathbb{E}(X^3) = p(p+1)(p+2)/\lambda^3$ *et* $m_4 = p(p+1)(p+2)(p+3)/\lambda^4$. *En développant le polynôme d'ordre 3 ou 4 dans l'expression de* μ_3 *et* μ_4 *et après simplification, on obtient*

$$\mu_3 = \frac{2p}{\lambda^3} \qquad et \qquad \mu_4 = \frac{3p^2 + 6p}{\lambda^4}$$

et l'expression de la variance limite

$$\Sigma^2 = \begin{pmatrix} p/\lambda^2 & 2p/\lambda^3 \\ 2p/\lambda^3 & (2p^2+6p)/\lambda^4 \end{pmatrix}.$$

En appliquant la δ-méthode à la fonction g telle que $(x,y) \mapsto (x^2/y, x/y)$ *differentiable sur* $\mathbb{R} \times \mathbb{R}_+^*$ *de Jacobienne*

$$Jg(x,y) = \begin{pmatrix} 2x/y & -x^2/y^2 \\ 1/y & -x/y^2 \end{pmatrix}.$$

D'où

$$Jg(p/\lambda, p/\lambda^2)\Sigma^2 Jg(p/\lambda, p/\lambda^2)^T = \begin{pmatrix} 2\lambda & -\lambda^2 \\ \lambda^2/p & -\lambda^3/p \end{pmatrix} \begin{pmatrix} p/\lambda^2 & 2p/\lambda^3 \\ 2p/\lambda^3 & (2p^2+6p)/\lambda^4 \end{pmatrix} \begin{pmatrix} 2\lambda & \lambda^2/p \\ -\lambda^2 & -\lambda^3/p \end{pmatrix}$$

et on obtient finalement

$$\sqrt{n}(T_n - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}_2\left(0_2, \begin{pmatrix} 2p(p+1) & 2\lambda(p+1) \\ 2\lambda(p+1) & \frac{\lambda^2}{p}(3+2p) \end{pmatrix}\right).$$

Sous les conditions de loi limite, tout Z-estimateur, et donc tout M-estimateur, est asymptotiquement normale. Comparer 2 tels estimateurs (si ils existent) revient donc à comparer leurs variances asymptotiques.

Théorème 5.4.2 *Sous les conditions de loi limite des Z-estimateurs, tout Z-estimateur T_n solution de*

$$\frac{1}{n}\sum_{i=1}^{n} \Phi(X_i, T_n) = 0_d$$

est moins efficace asymptotiquement que l'EMV (si il existe).

Démonstration : D'après le théorème de convergence des Z-estimateurs, on sait que T_n est asymptotiquement normal de variance asymptotique

$$V_\Phi(\theta) = (\mathbb{E}_\theta[J_\theta \Phi(X,\theta)])^{-1} \mathrm{Var}_\theta[\Phi(X,\theta)](\mathbb{E}_\theta[J_\theta \Phi(X,\theta)])^{-1^T}.$$

Or, dans le cas de l'EMV $\Phi = \nabla_\theta \log f(X,\theta)$ est le vecteur score. En utilisant les propriétés de celui-ci (cf. Chapitre 3) on obtient facilement

$$V_\Psi(\theta) = -(\mathbb{E}_\theta[\mathbb{H}_\theta \log f(X,\theta)])^{-1} = I(\theta)^{-1}.$$

L'EMV est donc asymptotiquement efficace. il suffit de prouver que $V_\Phi(\theta) \geq I^{-1}(\theta)$ pour tout $\theta \in \Theta$. Par définition d'un Z-estimateur, on a $\theta \mapsto \mathbb{E}_\theta(\Phi(X,\theta)) = 0_d$ comme fonction définie sur Θ. En dérivant terme à terme, on obtient

$$\begin{aligned} 0 &= \int_{\mathcal{X}} J_\theta(\Phi(x,\theta)) f(x,\theta) d\nu(x) + \int_{\mathcal{X}} \Phi(x,\theta) J_\theta f(x,\theta) d\nu(x) \\ &= \int_{\mathcal{X}} J_\theta(\Phi(x,\theta)) f(x,\theta)) d\nu(x) + \int_{\mathcal{X}} \Phi(x,\theta)[\nabla_\theta \log(f(x,\theta))]^T f(x,\theta) d\nu(x) \end{aligned}$$

autrement dit, en faisant apparaître le vecteur score $S(X,\theta)$

$$\mathbb{E}_\theta[J_\theta \Phi(X,\theta)] = -\mathbb{E}_\theta[\Phi(X,\theta) S(X,\theta)^T].$$

Pour simplifier les notations, on note par des majuscules les différents e.a. $J = J_\theta \Phi(X,\theta)$, $\Phi = \Phi(X,\theta)$ et $S = S(X,\theta)$ et on a obtenu $\mathbb{E}_\theta(J) = -\mathbb{E}_\theta(\Phi S^T)$. Par

passage au complémentaire, on a aussi $\mathbb{E}_\theta(J^T) = -\mathbb{E}_\theta(\Phi^T S)$. On rappelle qu'avec ces notations, $V_\Phi(\theta) = \mathbb{E}_\theta(J)^{-1}\mathrm{Var}_\theta(\Phi)(\mathbb{E}_\theta(J)^{-1})^T = \mathrm{Var}_\theta(\mathbb{E}_\theta(J)^{-1}\Phi)$. D'où

$$\begin{aligned}\mathrm{Var}_\theta(I^{-1}(\theta)S + \mathbb{E}_\theta(J)^{-1}\Phi) = I^{-1}(\theta)&\mathrm{Var}_\theta(S)I^{-1}(\theta)\\ &+ I^{-1}(\theta)\mathbb{E}_\theta(S\Phi^T)(\mathbb{E}_\theta(J)^{-1})^T + \mathbb{E}_\theta(J)^{-1}\mathbb{E}_\theta(\Phi S^T)I^{-1}(\theta) + V_\Phi(\theta).\end{aligned}$$

Par définition, $I^{-1}(\theta)\mathrm{Var}_\theta(S)I^{-1}(\theta) = I^{-1}(\theta)$ et d'après l'identité obtenue précédemment

$$I^{-1}(\theta)\mathbb{E}_\theta(S\Phi^T)(\mathbb{E}_\theta(J)^{-1})^T = \mathbb{E}_\theta(J)^{-1}\mathbb{E}_\theta(\Phi S^T)I^{-1}(\theta) = -I^{-1}(\theta)$$

d'où $\mathrm{Var}_\theta(I^{-1}(\theta)S + \mathbb{E}_\theta(J)^{-1}\Phi) = V_\Phi(\theta) - I^{-1}(\theta) \geq 0$ comme toute matrice de variance-covariance. □

Remarque 19 *La borne de Cramer-Rao asymptotique est la variance asymptotique minimale pour l'ensemble des Z- et M- estimateurs sous les conditions de loi limites. Il existe toutefois des estimateurs dont la variance asymptotique est plus petite que la borne de Cramer-Rao asymptotique. Ce ne sont pas des Z- ni des M-estimateurs.*

Chapitre 6

La racine de l'équation de vraisemblance

Dans le chapitre précédent, on a vu que, lorsqu'il existe, l'EMV est l'estimateur le plus efficace asymptotiquement parmi les M- et Z- estimateurs sous des conditions de loi limite sur la fonction de log-vraisemblance l_n. Pour étudier l'existence de l'EMV dans un modèle régulier, il est plus simple d'étudier l'existence de l'estimateur de la racine de l'équation de vraisemblance, appelée REV et noté $\hat{\theta}_n^{RV}$ ou plus simplement $\hat{\theta}_n$ (c.f. définition ci-dessous).

6.1 Conditions du premier et second ordre

Soit (P_θ, Θ) un modèle régulier. La fonction de log-vraisemblance l_n est bien définie par

$$l_n(\theta) = -\frac{1}{n}\sum_{i=1}^{n} \log f(X_i, \theta).$$

De plus, étant donné les conditions (H2) de dérivabilité satisfaites, l'EMV $\hat{\theta}_n^{MV} \in \Theta$ avec Θ ouvert est un Z-estimateur pour la fonction $\theta \mapsto \nabla_\theta l_n(\theta)$:

$$\nabla_\theta l_n(\hat{\theta}_n^{MV}) = -\frac{1}{n}\sum_{i=1}^{n} \nabla_\theta \log f(X_i, \hat{\theta}_n^{MV}) = 0 \tag{6.1}$$

Cette équation est appelée la condition du premier ordre ou condition nécéssaire ou équation de vraisemblance (EV).

L'EMV est aussi par définition un minimiseur local de la fonction de log-vraisemblance l_n. D'où la condition du second ordre qui est une condition suffi-

sante :

$$\mathbb{H}_\theta(l_n(\hat{\theta}_n^{MV})) = -\frac{1}{n}\sum_{i=1}^{n} \mathbb{H}_\theta \log f(X_i, \hat{\theta}_n^{MV}) > 0 \tag{6.2}$$

c'est à dire la matrice Hessienne de la fonction de log-vaisemblance est définie positive.

Définition 6.1.1 *L'estimateur de la racine de l'équation de vraisemblance ou REV, noté $\hat{\theta}_n$, est, si elle existe, une solution de l'EV.*

Remarque 20

1. *Si l'EMV existe, alors il coïncide avec une REV. L'existence de l'EMV et celle de la REV sont donc liée, la seconde étant plus facile à étudier car ayant lieu sous des conditions plus générales.*
2. *Si la REV existe et vérifie la condition du second ordre, alors il est un minimiseur local de la fonction de log-vraisemblance, i.e. un maximiseur local de la vraisemblance. Mais ce n'est pas forcément l'EMV (voir discussion ci-après).*
3. *L'hypothèse d'intégrabilité (H3) et la LFGN nous assure que*

$$-\frac{1}{n}\sum_{j=1}^{n} \nabla_\theta \log f(X_j, \theta) \xrightarrow{p.s.} -\mathbb{E}_\theta[\nabla_\theta \log f(X, \theta)] = -\mathbb{E}_\theta[S(X, \theta)] = 0,$$

$$-\frac{1}{n}\sum_{j=1}^{n} \mathbb{H}_\theta \log f(X_j, \theta) \xrightarrow{p.s.} -\mathbb{E}_\theta[\mathbb{H}_\theta \log f(X, \theta)] = I(\theta).$$

donc sous (H4) et pour n suffisamment grand, les conditions du premier et du second ordre ont des chances d'être réalisées dans un voisinage de la vraie valeur inconnue θ qui régit la loi P_θ dont est issu l'échantillon $(X_1, \ldots, X_n)$.

Le théorème suivant donne des conditions suffisantes pour que REV et EMV coïncident :

Théorème 6.1.1 *Si Θ est un intervalle ouvert de la forme $]\underline{\theta}, \overline{\theta}[$ pour $\underline{\theta}, \overline{\theta} \in (\mathbb{R} \cup \{\pm\infty\})^d$ alors une unique REV qui vérifie la condition du second ordre coïncide avec l'unique EMV.*

Démonstration : Comme la REV $\hat{\theta}_n$ vérifie la condition du second ordre, elle réalise un minimum local de l_n. Montrons que si Θ est un intervalle alors c'est un maximum global. La fonction $\theta \mapsto \nabla_\theta l_n(\theta)$ s'annule en un unique point $\hat{\theta}_n$ de Θ. C'est une fonction continue donc elle est de signe constant de par et d'autre de $\hat{\theta}_n$

sur l'intervalle Θ. Autrement dit $\hat{\theta}_n$ est un extremum global de l_n. Mais c'est aussi un minimum local, donc c'est un minimal global et donc un EMV. Enfin, si il y avait un autre EMV, ce serait aussi une REV distinct ce qui est en contradiction avec l'énoncé donc l'EMV est unique. □

6.2 Propriétés non asymptotiques de la REV

6.2.1 Exhaustivité et reparamétrisation

On s'intéresse ici aux propriétés non asymptotiques de l'EMV. Dans un modèle paramétrique (P_θ, Θ) pas forcément régulier on suppose que l'EMV $\hat{\theta}_n^{MV}$ de θ existe et est unique : c'est l'unique maximiseur global de la vraisemblance L_n.

Théorème 6.2.1 *Si T_n est une statistique exhaustive pour θ alors l'EMV $\hat{\theta}_n^{MV}$ est une fonction de T_n.*

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

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

L'EMV $\hat{\theta}_n^{MV}$ satisfait par définition $L_n(\hat{\theta}_n^{MV}) \geq L_n(\theta)$ soit $g(T_n, \hat{\theta}_n^{MV}) \geq g(T_n, \theta)$ pour tout $\theta \in \Theta$. Comme tout estimateur, $\hat{\theta}_n^{MV}$ ne doit pas dépendre de θ et que le critère à maximiser ne dépend que de T_n et θ, $\hat{\theta}_n^{MV}$ est forcément fonction de T_n. □

Remarque 21 *l'EMV lui-même n'est pas forcément une statistique exhaustif.*

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

Théorème 6.2.2 *(Théorème de Zehna) Pour n'importe quelle application φ de Θ dans Θ, si $\hat{\theta}_n^{MV}$ est unique alors l'estimateur $\varphi(\hat{\theta}_n^{MV})$ est un EMV de $\varphi(\theta)$.*

Démonstration : On définit pour tout $\eta \in \varphi(\Theta)$ la fonction de vraisemblance de η :

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

On a supposé que l'EMV était le maximiseur global de L_n d'où

$$\sup_{\eta \in \varphi(\Theta)} L_n^*(\eta) = \sup_{\theta \in \Theta} L_n(\theta) = L_n(\hat{\theta}_n^{MV}).$$

Or,

$$L_n(\hat{\theta}_n^{MV}) = \sup_{\theta\in\Theta \ / \ \varphi(\theta)=\varphi(\hat{\theta}_n^{MV})} L_n(\theta) = L_n^*(\varphi(\hat{\theta}_n^{MV})).$$

Il vient que $\sup_{\eta\in\varphi(\Theta)} L_n^*(\eta) = L_n^*(\varphi(\hat{\theta}_n))$ et donc $\varphi(\hat{\theta}_n)$ (qui est clairement dans $\varphi(\Theta)$) est bien un maximiseur de L_n^* (pas nécéssairement unique). C'est donc un EMV de $\varphi(\theta)$. □

6.2.2 Cas d'un modèle de la famille exponentielle

Considérons un modèle de la famille exponentielle identifiable (P_θ, Θ) :

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

Si $\alpha \in C^2$ et $\mathbb{E}_\theta(T_j(X)^2) < \infty$ pour tout $1 \le j \le r$ alors le modèle est régulier et on vérifie automatiquement que la fonction $\theta \mapsto \log c(\theta) \in C^2$. On peut donc calculer la REV $\hat{\theta}_n$ de θ, solution de la condition du premier ordre :

Théorème 6.2.3 *Soit (P_θ, Θ) un modèle de la famille exponentielle identifiable et tel que $\mathbb{E}_\theta(T_j(X)^2) < \infty$ pour tout $1 \le j \le r$. Alors si la REV $\hat{\theta}_n$ de θ existe et qu'elle est sans biais alors elle coïncide aussi avec l'estimateur de variance minimale.*

Démonstration : On écrit l'EV dans le cas de la famille exponentielle :

$$\begin{aligned}
L_n(\theta) &= \prod_{i=1}^{n} h(X_i)\exp\left(\sum_{i=1}^{n}\sum_{j=1}^{r}\alpha_j(\theta)T_j(X_i) + n\log c(\theta)\right) \\
l_n(\theta) &= cste - \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{r}\alpha_j(\theta)T_j(X_i) - \log c(\theta) \\
\nabla_\theta l_n(W_n) &= 0 \Leftrightarrow \frac{\partial \log c(W_n)}{\partial\theta_k} = -\frac{1}{n}\sum_{j=1}^{n}\frac{\partial\alpha_j(\theta)}{\partial\theta_k}\sum_{i=1}^{n}T_j(X_i) \qquad 1 \le k \le r.
\end{aligned}$$

Cette équation ne dépend que de la statistique exhaustive complète

$$S_n = \left(\sum_{i=1}^{n}T_1(X_i), \ldots, \sum_{i=1}^{n}T_r(X_i)\right)$$

donc si la REV existe elle s'exprime en fonction de S_n et la dernière propriété découle du théorème de Lehman-Scheffé. □

Exemple 6.2.1

- *Soit* $(\mathcal{N}(\mu,\sigma^2), \theta = \mu \in \mathbb{R})$ *alors l'EV fournit une unique REV* $\hat{\theta}_n = \overline{X}_n$ *qui est aussi l'unique EMV car* Θ *est un intervalle et* $l_n''(\hat{\theta}_n) = 1/(2\sigma^2) > 0$. *De plus, il est sans biais, c'est donc aussi l'estimateur sans biais de variance minimale (unique d'après le théorème de Lehman-Scheffé).*
- *Soit* $(\mathcal{N}(\mu,\sigma^2), \theta = (\mu,\sigma^2) \in \mathbb{R}\times\mathbb{R}_+^*)$ *alors l'EV correspond au système*

$$\begin{cases} \dfrac{\hat{\mu}-\overline{X}_n}{2\hat{\sigma}^2} & =0 \\ \dfrac{1}{2}\left(\dfrac{1}{\hat{\sigma}^2} - \dfrac{(\overline{X}_n-\hat{\mu})^2}{\hat{\sigma}^4}\right) & =0 \end{cases}$$

d'où l'unique REV $\hat{\theta}_n = (\hat{\mu},\hat{\sigma}^2) = (\overline{X}_n, S_n^2)$ *qui est aussi l'unique EMV car* Θ *est un intervalle ouvert et la condition du second ordre est satisfaite :*

$$\mathbf{H}_\theta(l_n(\hat{\theta}_n)) = \begin{pmatrix} \frac{1}{2\hat{\sigma}^2} & 0 \\ 0 & \frac{1}{2\hat{\sigma}^4} \end{pmatrix} > 0.$$

Par contre, S_n^2 *étant biaisé,* $\hat{\theta}_n$ *est biaisé et ne coïncide pas avec l'estimateur sans biais de variance minimale* $(\overline{X}_n, S_n^{2\prime})$.

6.3 Propriétés asymptotiques de la REV

Soit (P_θ,Θ) un modèle régulier qui n'appartient pas forcément à la famille exponentielle. Dans ce contexte, il faut faire appel à l'asymptotique pour étudier l'existence de la REV $\hat{\theta}_n$ et en déduire celle de l'EMV.

Théorème 6.3.1 *Soit* (P_θ,Θ) *un modèle régulier identifiable, i.e. la fonction* $\theta \mapsto P_\theta$ *est injective, alors à partir d'un certain rang il existe une suite de REV* $\hat{\theta}_n$ *qui est asymptotiquement efficace :*

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}_d(0_d, I^{-1}(\theta)).$$

Démonstration : La preuve se décompose en 2 parties. Dans la première partie, on va montrer qu'à partir d'un certain rang il existe une suite de REV convergeant vers θ. On montrera que celle-ci est alors nécessairement asymptotiquement efficace dans un deuxième temps.

On note $\psi_n(a) = \log(L_n(a)/L_n(\theta))$ telle que

$$\nabla_\theta l_n(\theta) = 0 \Leftrightarrow \nabla_\theta \psi_n(\theta) = 0.$$

D'après la loi forte des grands nombres

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

ou $-K(\theta,a) \leq 0$ d'après l'inégalité de Jensen appliquée à la fonction convexe $x \mapsto -\log(x)$. De plus, on montre que $K(\theta,a)= 0$ si et seulement si $f(X,a) = f(X,\theta)$ avec probabilité 1, soit, en utilisant l'identifiabilité du modèle, $\theta = a$. Donc pour tout $\varepsilon > 0$ suffisamment petit pour que $[\theta-\varepsilon;\theta+\varepsilon]^d \subseteq \Theta$, il existe N_ε tel qu'on ait :

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

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

$$P_\theta\left(\forall\ n > N_\varepsilon,\ \exists \text{ une REV } \ \hat{\theta}_n \text{ et } \ \|\hat{\theta}_n - \theta\| < \epsilon\right) = 1.$$

En particulier, dès que $[\theta_0-\varepsilon;\theta_0+\varepsilon]^d \subseteq \Theta$ (toujours possible car Θ est ouvert) on a

$$P_\theta\left(\forall\ n > N_\varepsilon,\ \exists \text{ une REV } \ \hat{\theta}_n\right) = 1,$$

donc à partir du rang N_ε il existe une suite de REV $\hat{\theta}_n$. Remarquons que par construction cette suite de REV, étant choisi comme étant la plus proche de θ, ne dépend pas de ε (seul le rang N_ε dépend de ε). Donc pour tout $\epsilon > 0$ on a en particulier

$$\lim_{n\to\infty} P_\theta\left(\|\hat{\theta}_n - \theta\| < \epsilon\right) = 1$$

(la suite est même égale à 1 à partir du rang N_ε). Donc à partir d'un certain rang il existe bien une suite de REV $\hat{\theta}_n$ qui converge vers θ.

Montrons que cette suite de REV $\hat{\theta}_n$ convergente est aussi asymptotiquement efficace pour θ. On définit pour tout $a \in \Theta$ la fonction

$$\varphi_n(a) = \frac{\sum_{j=1}^n S(X_j,a)}{n} = \frac{1}{n}\sum_{j=1}^{n} \nabla_\theta(\log f)(X_j,a).$$

Soit $1 \leq j \leq d$ un indice quelconque. D'après le développement de Taylor à l'ordre 1 de la fonction φ_n au point θ, il existe $\overline{\theta}_n = (\overline{\theta}_{n,i})'_{1\leq i\leq d}$ vérifiant

$$0 = \varphi_{n,j}(\hat{\theta}_n) = \varphi_{n,j}(\theta) + \nabla(\varphi_{n,j})(\overline{\theta}_n)^T(\hat{\theta}_n - \theta) \text{ et } \overline{\theta}_{n,i} \in [\min(\theta_i, \hat{\theta}_{n,i}), \max(\theta_i, \hat{\theta}_{n,i})],$$

soit

$$\left(I_j(\theta) - I_j(\theta) - \nabla\varphi_{n,j}(\overline{\theta}_n)\right)^T (\hat{\theta}_n - \theta) = \varphi_{n,j}(\theta)$$

où I_j est le j-ème vecteur colonne de I. On sait que l'ensemble $\mathcal{C}(K)$ des fonctions continues définies sur le compact $K = \{a \in \Theta;\ \|a - \theta\| \leq \varepsilon\}$ et muni de la norme uniforme $\|\cdot\|_K$ est un espace de Banach. Sous (H3), on vérifie pour tout$1 \leq i, j \leq d$ que

$$\mathbb{E}_\theta \|\partial^2 \log f(x, a)/\partial\theta_i\theta_j\|_K < \infty.$$

Si ∇_i est la dérivée par rapport à la i-ème coordonnée θ_i et $I_{i,j}$ et le coefficient i, j de l'information de Fisher, en appliquant la LFGN on obtient

$$P_\theta\left(\lim_{n\to\infty} \|\nabla_i\varphi_{n,j}(a) + I_{i,j}(a)\|_K = 0\right) = 1.$$

Mais la convergence uniforme sur K et la continuité de $a \to \nabla_i\varphi_{n,j}(a)$ entraîne la continuité de $a \to I_{i,j}(a)$ sur K.

Par hypothèse, on sait que $\overline{\theta}_n \xrightarrow{P} \theta$. En particulier, $\|\nabla_i\varphi_{n,j}(\overline{\theta}_n) + I_{i,j}(\theta)\| \leq \|\nabla_i\varphi_{n,j}(\overline{\theta}_n) + I_{i,j}(\hat{\theta}_n)\| + \varepsilon_n$ avec $\varepsilon_n \xrightarrow{P} 0$ donc

$$\begin{aligned} \|\nabla_i\varphi_{n,j}(\overline{\theta}_n) + I_{i,j}(\theta)\| &\leq \|\nabla_i\varphi_{n,j}(\overline{\theta}_n) - I_{i,j}(\overline{\theta}_n)\| + \varepsilon_n \\ &\leq \|\nabla_i\varphi_{n,j}(a) - I_{i,j}(a)\|_K + \varepsilon_n \xrightarrow{P} 0. \end{aligned}$$

Il s'en suit que pour tout $1 \leq j \leq d$ on a

$$(I_j(\theta) + o_P(1))^T (\hat{\theta}_n - \theta) = \varphi_{n,j}(\theta_0)$$

où $o_P(1)$ est un terme aléatoire qui tend vers 0 en probabilité. En écrivant la forme vectorielle de ce système d'équation et en multipliant par $\sqrt{n}$, on trouve

$$\sqrt{n}(I(\theta) + o_{\mathbb{P}}(1))(\hat{\theta}_n - \theta) = \sqrt{n}\varphi_n(\theta)$$

On conclut par le TCL appliqué aux vecteurs scores qui donne $\sqrt{n}\varphi_n(\theta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(0, I(\theta_0))$, le théorème de Slutsky et la δ-méthode sous (H4). □

Remarque 22 *On vient de construire une suite de REV asymptotiquement efficace comme étant le maximum local le plus proche de θ. La construction théorique de cette suite fait donc appel à la connaissance de θ inconnu ! En pratique on ne peut donc pas utiliser cette suite de REV, d'où le corollaire suivant.*

Corollaire 6.3.1 *Soit (P_θ, Θ) un modèle régulier identifiable. Si, à partir d'un certain rang, il existe une unique REV $\hat{\theta}_n$ alors elle est asymptotiquement efficace.*

Démonstration : Dans la preuve du théorème 6.3.1 on construit une suite de REV qui est asymptotiquement efficace. Comme la REV $\hat{\theta}_n$ est supposée unique, elle coïncide nécessairement avec celle construite précédemment et est donc asymptotiquement efficace. □

Exemple 6.3.1 *Soit le modèle régulier $(\mathcal{E}(\theta), \theta > 0)$ de vraisemblance*

$$L_n(\theta) = \theta^n e^{-\theta \sum_{j=1}^n X_j}$$

et de fonction de log-vraisemblance

$$l_n(\theta) = \frac{1}{n}\theta \sum_{j=1}^n X_j - \log \theta.$$

On trouve une unique REV $\hat{\theta}_n = (\overline{X}_n)^{-1}$. Soit on vérifie que le modèle régulier et identifiable, et d'après le corollaire 6.3.1 cette suite coïncide avec l'EMV car $\Theta =]0, +\infty[$ et la condition du second ordre est satisfaite et comme $I(\theta) = \theta^{-2}$ on obtient l'efficacité asymptotique de l'EMV

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \theta^2).$$

Soit on utilise le TLC et la δ-méthode pour obtenir la normalité asymptotique

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \theta^2)$$

et on remarque que la variance asymptotique atteint la borne de Cramer-Rao asymptotique.

6.4 Conclusion sur l'estimation ponctuelle

En pratique, supposons que la modélisation de l'expérience renouvelable fournit un modèle régulier identifiable (P_θ, Θ) tel que $\theta \mapsto P_\theta$ soit connu et tel que l'échantillon observé $(X_1, \ldots, X_n)$ soit issu de la loi P_θ où θ est le paramètre d'intérêt inconnu. On écrit alors l'équation de la vraisemblance.

1. Si on obtient l'expression d'une racine de ce système, alors on vérifie que cette racine coïncide bien avec le maximum de la vraisemblance et on calcule le biais et la variance de cet estimateur.

   (a) Si l'EMV est biaisé, on le corrige pour obtenir un estimateur sans biais puis calculer la variance de cet estimateur sans biais. Si l'estimateur corrigé à une variance plus petite que la somme du biais au carré et de la variance de l'EMV, on le préfère à l'EMV.
      i. Si on est dans un modèle de la famille exponentielle c'est l'estimateur de variance minimale.
      ii. Si on n'est pas dans un modèle de la famille exponentielle on compare sa variance avec la borne de Cramer-Rao pour voir si il n'est pas efficace.
   (b) Si l'EMV est sans biais, on reprend les points i. et ii.
2. Si on n'obtient pas l'expression de la REV, alors on essaie la méthode des moments (généralisés ou non). Il faut vérifier que l'estimateur obtenu est asymptotiquement normal et comparer sa variance asymptotique avec la borne de Cramer-Rao asymptotique. Si l'estimateur est asymptotiquement efficace, on reprend les points (a) et (b).

Exemple 6.4.1 *Soit le modèle Gamma* $(\gamma(p,\lambda), \theta = (p,\lambda) \in]0,\infty[^2)$*, on peut vérifier que c'est un modèle régulier de la famille exponentielle. la fonction de log-vraisemblance vaut*

$$l_n(\theta) = \lambda \sum_{i=1}^{n} X_i - \sum_{i=1}^{n} (p-1)\log(X_i) - np\log(\lambda) + n\log(\Gamma(p))$$

et l'équation de vraisemblance est le système :

$$\begin{cases} -\sum_{i=1}^{n} \log(X_i) - \log(\lambda) + n\Gamma'(p)/\Gamma(p) = 0 \\ \sum_{i=1}^{n} X_i - np/\lambda = 0 \end{cases}$$

On peut vérifier que ce système admet une unique solution qui est l'EMV. Toutefois, celle-ci n'a pas d'expression analytique car l'inverse de la fonction Gamma n'en a pas. On utilise donc l'estimateur obtenu par la méthode des moments

$$T_n = \left(\frac{(\overline{X}_n)^2}{S_n^2}, \frac{\overline{X}_n}{S_n^2} \right)$$

qui est asymptotiquement normal

$$\sqrt{n}(T_n - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}_2 \left(0_2, \begin{pmatrix} 2p(p+1) & 2\lambda(p+1) \\ 2\lambda(p+1) & \frac{\lambda^2}{p}(3+2p) \end{pmatrix} \right).$$

D'après le théorème de comparaison des M- et Z- estimateurs, on sait que cette variance asymptotique est plus grande que la Borne de Cramer Rao asymptotique $I^{-1}(\theta)$.

Chapitre 7

Régions de confiance

On se place comme précédemment dans le cadre d'un modèle paramétrique (P_θ, Θ) où le paramètre d'intérêt θ est inconnu. A partir de l'échantillon $(X_1, \ldots, X_n)$ issu de la loi P_θ on veut inférer sur le paramètre θ. Plutôt que de donner une valeur approximative de θ, on cherche désormais à trouver un sous-ensemble de Θ dans lequel le paramètre inconnu θ a une forte probabilité d'appartenir. Pour cela, on utilise une statistique $T_n \in \mathcal{Y}$ qui n'est plus un estimateur $\mathcal{Y} \neq \Theta$ mais un sous ensemble de $\Theta : \mathcal{Y} = \mathcal{P}(\Theta)$. On notera $T_n = C_n$ pour différencier avec l'estimation ponctuelle.

7.1 Définition

Soit $0 < \alpha < 1$ un niveau de risque fixé par le statisticien.

Définition 7.1.1 *La statistique $C_n \in \mathcal{Y} = \mathcal{P}(\Theta)$ est une région de confiance de niveau (de confiance) $1 - \alpha$ pour θ si elle ne dépend pas du paramètre inconnu θ et si,*

$$P_\theta(\theta \in C_n) \geq 1 - \alpha, \quad \text{pour tout } \theta \in \Theta.$$

La statistique C_n est une région de confiance de taille $1 - \alpha$ pour θ lorsque

$$P_\theta(\theta \in C_n) = 1 - \alpha, \quad \text{pour tout } \theta \in \Theta.$$

Remarque 23

- *Par passage au complémentaire, le niveau de risque α correspond à une majoration de la probabilité que le vrai paramètre θ ne soit pas dans C_n.*
- *La région de confiance C_n dépend de α qui est connu par le statisticien, c'est lui qui fixe le niveau de risque.*

La région de confiance C_n est une statistique non paramétrique car l'ensemble des sous ensembles de Θ noté $\mathcal{P}(\Theta)$ est de dimension infinie (hormis le cas où Θ est fini). On se ramène à une statistique paramétrique en ne considérant que des sous ensembles de forme particulière de Θ :

- dans le cas unidimensionnel $\Theta \subseteq \mathbb{R}$ ($d = 1$), on choisit C_n de la forme $C_n = [A_n, B_n]$ où A_n et B_n sont deux estimateurs de θ vérifiant $A_n \leq B_n$. La région de confiance obtenue est appelée intervalle de confiance.
- dans le cas multidimensionnel $\Theta \subseteq \mathbb{R}^d$ avec $d \geq 1$, on choisit C_n de la forme d'une ellipsoïde :

$$C_n = \{a \in \Theta \ / \ \|P_n(a - W_n)\|^2 \leq M_n\}$$

où W_n est un estimateur de θ, P_n est une matrice aléatoire ne dépendant pas de θ correspondant à un changement de base et $M_n \geq 0$ pour $1 \leq i \leq d$ est une statistique ne dépendant pas de θ donnant la largeur de la région de confiance dans cette nouvelle base.

Remarque 24 *Dans le cas unidimensionnelle, un intervalle de confiance est centré en l'estimateur W_n lorsqu'il est déterminé par les relations $A_n = W_n - M_n$ et $B_n = W_n + M_n$ avec $M_n > 0$.*

Avant d'étudier la construction de telles régions de confiance, nous en donnons un exemple connu :

Exemple 7.1.1 *Soit le modèle Gaussien $(\mathcal{N}(\mu, \sigma^2), \theta = \mu \in \mathbb{R})$ avec $\sigma > 0$ connu. On considère alors l'intervalle de confiance centré en $\overline{X}_n$ de largeur $M_n = \sigma q^N_{1-\alpha/2}/\sqrt{n} > 0$:*

$$C_n = [\overline{X}_n - \frac{\sigma}{\sqrt{n}} q^N_{1-\alpha/2}, \overline{X}_n + \frac{\sigma}{\sqrt{n}} q^N_{1-\alpha/2}]$$

où $q^N_{1-\alpha/2}$ est le quantile d'ordre $\alpha/2$ de la loi normale centrée réduite, i.e. $F_N(q^N_{1-\alpha/2}) = 1 - \alpha/2$ avec $N \sim \mathcal{N}(0, 1)$.

7.2 Fonctions pivotales

Pour construire une région de confiance de taille $1-\alpha$, on utilise une fonction pivotale (réelle) :

Définition 7.2.1 *La fonction aléatoire $Q_n(\theta) \in \mathbb{R}$ du paramètre θ dont on veut déterminer une région de confiance est une fonction pivotale si c'est une fonction de θ dont la loi est connue.*

Une fonction pivotale n'est pas unique, en pratique on utilise la fonction pivotale la plus simple possible.

Exemple 7.2.1 *Dans le cas du modèle Gaussien multidimensionnel* $(\mathcal{N}_d(\theta, \Sigma^2), \theta \in \mathbb{R}^d)$ *avec* Σ^2 *connu, d'après le théorème de Cochran la statistique*

$$Q_n(\theta) = n\|\Sigma^{-1}(\overline{X}_n - \theta)\|^2 \sim \chi^2_d$$

est une fonction pivotale dite du χ^2*.*

Une fois la fonction pivotale réelle obtenue et $0 < \alpha < 1$ fixé on détermine la région de confiance de taille $1 - \alpha$ grâce à la proposition suivante :

Proposition 7.2.1 *Soit* $Q_n(\theta)$ *une fonction pivotale dont la loi est absolument continue et soit* $0 < \alpha < 1$ *fixé. Alors pour tout* $0 \leq \gamma \leq \alpha$ *on obtient une région de confiance de taille* $1 - \alpha$ *de la forme*

$$C_n(\gamma) = S_n^{-1}([q_\gamma^{Q_n}, q_{1-\alpha+\gamma}^{Q_n}]) = \{a \in \Theta \ / \ q_\gamma^{Q_n} \leq S_n(a) \leq q_{1-\alpha+\gamma}^{Q_n}\},$$

où $q_\gamma^{Q_n}$ *est le quantile d'ordre* γ *de la loi de* $Q_n(\theta)$ *:*

$$P_\theta(Q_n(\theta) \leq q_\gamma^{Q_n}) = \gamma.$$

Par définition de la fonction pivotale, la loi de $Q_n(\theta)$ ne dépend pas de θ donc $q_\gamma^{S_n}$ est bien définie : le quantile ne dépend pas non plus de θ (qui reste inconnu).
Démonstration : Il suffit de vérifier que

$$P_\theta(\theta \in C_n(\gamma)) = 1 - \alpha$$

soit par passage au complémentaire

$$\begin{aligned} P_\theta(\theta \notin C_n(\gamma)) = \alpha = P_\theta(Q_n(\theta) \notin [q_\gamma^{Q_n}, q_{1-\alpha+\gamma}^{Q_n}]) \\ = P_\theta(Q_n(\theta) < q_\gamma^{Q_n}) + 1 - P_\theta(Q_n(\theta) \leq q_{1-\alpha+\gamma}^{Q_n}) \end{aligned}$$

par définition de C_n et par définition des quantiles. □

Remarque 25

- *Pour chaque* γ *choisit correspond une région de confiance* $C_n(\gamma)$*. En théorie, il faut choisir* γ *qui correspond à la région* $C_n(\gamma)$ *d'aire la plus petite possible. En pratique, si la loi de la fonction pivotale est presque symétrique par rapport à son axe modal (la verticale passant par son mode) et par symétrie on choisit* $\gamma = \alpha/2$*. Sinon on choisit* $\gamma = 0$ *ou* $\gamma = \alpha$ *pour simplifier l'expression de la région de confiance en comparant les aires de* $C_n(0)$ *et* $C_n(\alpha)$*. Un loi du* χ^2_k *est presque symétrique par rapport à son axe modale si* k *est grand* ($k \propto n$) *et ne l'est plus si* k *est petit* ($k \propto d$)*.*

– *Dans le cas où la fonction pivotale $Q_n(\theta)$ est réelle discrète, alors on ne peut pas systématiquement obtenir des régions de confiance de taille $1-\alpha$ car par définition des quantiles on peut avoir $P_\theta(Q_n(\theta) < q_\gamma^{S_n}) \neq \gamma$. Par contre, avec un procédé similaire il est toujours possible de trouver une région de confiance de niveau $1-\alpha$ même dans ce cas.*

Exemple 7.2.2

– *Dans le cas Gaussien $(\mathcal{N}(\mu, \sigma^2), \theta = \mu \in \mathbb{R})$ avec $\sigma^2 > 0$ connu, la fonction pivotale vaut*

$$Q_n(\theta) = \sqrt{n}\sigma^{-1}(\overline{X}_n - \theta) \sim \mathcal{N}(0,1).$$

Par symétrie de la loi normale, on choisit $\gamma = \alpha/2$ d'où l'intervalle de confiance centré

$$C_n = [\overline{X}_n - \frac{\sigma}{\sqrt{n}} q^N_{1-\alpha/2}, \overline{X}_n + \frac{\sigma}{\sqrt{n}} q^N_{1-\alpha/2}].$$

– *Dans le cas Gaussien multidimensionnel $\mathcal{N}_d(\theta, \Sigma^2), \theta \in \mathbb{R}^d)$ avec Σ^2 connu, on a*

$$Q_n(\theta) = n\|\Sigma^{-1}(\overline{X}_n - \theta)\|^2 \sim \chi^2_d.$$

Dans un cas multidimensionnel comme celui-ci, on choisit $\gamma = 0$ de manière à simplifier l'expression de C_n car $q_\gamma^{S_n} = q_0^{\chi^2_d} = 0$, d'où la région de confiance centrée en $\overline{X}_n$

$$C_n = \{a \in \mathbb{R}^d \ / \ n(\overline{X}_n - a)^T \Sigma^{2^{-1}} (\overline{X}_n - a) \leq q^{\chi^2_d}_{1-\alpha}\}.$$

– *Dans le cas Gaussien unidimensionnel $(\mathcal{N}(\mu, \sigma^2), \theta = (\mu, \sigma^2 \in \mathbb{R})$ avec $\sigma^2 > 0$ connu, on peut aussi utiliser la fonction pivotale*

$$\frac{n(\overline{X}_n - a)^2}{\sigma^2} \sim \chi^2_1$$

et choisir, comme la loi χ^2_1 n'est pas symétrique par rapport à son axe modal, l'intervalle de confiance

$$C_n = \{a \in \Theta \ / \ n(\overline{X}_n - a)^2 \leq \sigma^2 q^{\chi^2_1}_{1-\alpha}\} = [\overline{X}_n - \frac{\sigma}{\sqrt{n}} \sqrt{q^{\chi^2_1}_{1-\alpha}}, \overline{X}_n + \frac{\sigma}{\sqrt{n}} \sqrt{q^{\chi^2_1}_{1-\alpha}}].$$

Cet intervalle de confiance étant centré en $\overline{X}_n$ et de taille $1-\alpha$, c'est le même que celui précédemment obtenu. On peut effectivement vérifier que $q^N_{1-\alpha/2} = \sqrt{q^{\chi^2_1}_{1-\alpha}}$ car $N^2 = \chi^2_1$.

– *Soit le modèle exponentiel* $(\mathcal{E}(\theta), \theta > 0)$ *alors* $Y = 2X\theta \sim \chi^2_2$. *D'où*

$$Q_n(\theta) = 2\theta \sum_{i=1}^{n} X_i = 2n\theta \overline{X}_n \sim \chi^2_{2n}.$$

Pour n *grand* ($n \geq 50$), *comme* $\sum_{i=1}^{2n} N_i^2 \sim \chi^2_{2n}$ *avec* $N_i \sim \mathcal{N}(0,1)$, *l'approximation normale pour la somme partielle à lieu*

$$\chi^2_{2n} \overset{\mathcal{L}}{\approx} \mathcal{N}(2n, 4n).$$

Lorsque l'approximation normale a lieu, la loi est quasi-symétrique par rapport à l'axe modal car toute loi normale est symétrique par rapport à l'axe modal. On choisit donc $\gamma = \alpha/2$ *et comme*

$$P_\theta(q^{\chi^2_{2n}}_{\alpha/2} \leq 2n\theta\overline{X}_n \leq q^{\chi^2_{2n}}_{1-\alpha/2}) = 1 - \alpha$$

et on en déduit l'intervalle de confiance de taille $1 - \alpha$:

$$\left[\frac{q^{\chi^2_{2n}}_{\alpha/2}}{2n\overline{X}_n}, \frac{q^{\chi^2_{2n}}_{1-\alpha/2}}{2n\overline{X}_n} \right].$$

– *Dans le cas Gaussien unidimensionnel* $(\mathcal{N}(\mu, \sigma^2), \theta = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}^*_+)$ *on peut utiliser la fonction pivotale de Student*

$$\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sqrt{S_n^{2\prime}}} \sim t_{n-1}$$

où t_{n-1} *est la loi de Student à* $n-1$ *degrés de liberté. Comme la loi de Student est symétrique par rapport à son axe modal, on a une RC de taille* $1 - \alpha$

$$C_n = \left[\overline{X}_n - \frac{\sqrt{S_n^{2\prime}}}{\sqrt{n}} q^{t_{n-1}}_{1-\alpha/2}, \overline{X}_n + \frac{\sqrt{S_n^{2\prime}}}{\sqrt{n}} q^{t_{n-1}}_{1-\alpha/2} \right] \times \mathbb{R}^*_+.$$

Cette RC est d'aire infini mais l'IC correspondant sur μ *de niveau* $1-\alpha$ *est de longueur fini. Elle donne donc un encadrement précis du paramètre* μ *inconnu ceci indépendamment de la valeur de* σ^2. *On parle d'IC sur* μ *avec* σ^2 *inconnu de taille* $1 - \alpha$.

7.3 Régions de confiance asymptotiques

Il n'est pas toujours possible de construire une région de confiance de taille fixé lorsqu'aucune fonction pivotale n'est pas connue.

Exemple 7.3.1 *Soit le modèle* $(P_\theta, \theta = \mathbb{E}_\theta(X) \in \mathbb{R})$ *avec* $\sigma^2 > 0$ *connu. Alors* $\sqrt{n}\sigma^{-1}(\overline{X}_n - \theta)$ *n'est pas une fonction pivotale car la forme de* P_θ *n'étant pas spécifiée on ne connaît pas la loi de* $\overline{X}_n$.

On utilise alors une fonction pivotale asymptotique :

Définition 7.3.1 *La fonction aléatoire* $Q_n(\theta) \in \mathbb{R}$ *du paramètre* θ *dont on veut déterminer une région de confiance est une fonction pivotale asymptotique si c'est une fonction de* θ *dont la loi limite ne dépend pas du paramètre inconnu* $\theta \in \Theta$, *i.e.* $Q_n(\theta) \xrightarrow{\mathcal{L}} Y$ *où la loi de* Y *est connue.*

Exemple 7.3.2 *Soit le modèle* $(P_\theta, \theta = \mathbb{E}_\theta(X) \in \mathbb{R})$ *avec* $Var_\theta(X) = \sigma^2 > 0$ *connu. Alors* $\sqrt{n}\sigma^{-1}(\overline{X}_n - \theta)$ *est une fonction pivotale asymptotique de loi limite* $\mathcal{N}(0,1)$ *d'après le TCL.*

En suivant le schéma de construction de la section précédente, on obtient alors des régions de confiances de taille asymptotique $1-\alpha$:

Définition 7.3.2 *La statistique* C_n *est une RC de niveau (taille) asymptotique* $1-\alpha$ *pour* θ *lorsque*

$$\lim_{n\to\infty} P_\theta(\theta \in C_n) \leq \ (=)\ 1-\alpha, \qquad \text{pour tout } \theta \in \Theta.$$

Exemple 7.3.3 *Soit le modèle* $(P_\theta, \theta = \mathbb{E}_\theta(X) \in \mathbb{R})$ *avec* $Var_\theta(X) = \sigma^2 > 0$ *connu. Alors* $\sqrt{n}\sigma^{-1}(\overline{X}_n - \theta)$ *est une fonction pivotale asymptotique de loi limite* $\mathcal{N}(0,1)$ *d'après le TCL. On en déduit* C_n *l'intervalle de confiance centré en* $\overline{X}_n$ *de taille asymptotique* $1-\alpha$ *de la forme*

$$C_n = [\overline{X}_n - \frac{\sigma}{\sqrt{n}} q^N_{1-\alpha/2}, \overline{X}_n + \frac{\sigma}{\sqrt{n}} q^N_{1-\alpha/2}].$$

7.4 Fonctions pivotales asymptotiques usuelles

Dans un modèle paramétriques (P_θ, Θ) l'existence d'une fonction pivotale asymptotiques découle de l'existence d'un estimateur T_n asymptotiquement normal.

Théorème 7.4.1 *Soit un modèle paramétrique* (P_θ, Θ) *pour lequel il existe un estimateur asymptotiquement normal*

$$\sqrt{n}(T_n - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}_d(0_d, V(\theta)) \qquad \forall \theta \in \Theta$$

où $V(\theta)$ *est la variance asymptotique lorsque* $X_1, \ldots, X_n \sim P_\theta$. *On suppose connue la fonction* V *continue de* Θ *dans l'ensemble des matrices* $d \times d$ *symétrique définie positive. Alors*

$$n(T_n - \theta)^T V^{-1}(\theta)(T_n - \theta) \qquad et \qquad n(T_n - \theta)^T V^{-1}(T_n)(T_n - \theta)$$

sont des fonctions pivotales réelles pour θ *de loi limite une* χ^2_d.

Démonstration : Une application du théorème de Cochrane nous montre que la loi limite de $n(T_n-\theta)^T V^{-1}(\theta)(T_n-\theta) = \|\sqrt{n}V^{-1/2}(\theta)(T_n-\theta)\|^2$ est une χ^2_d qui ne dépend pas de θ. Pour la seconde qui vaut $\sqrt{n}\|V^{-1/2}(T_n)(T_n-\theta)\|^2$, il suffit de remarquer que $T_n \xrightarrow{P} \theta$ comme tout estimateur asymptotiquement normal et donc $V(T_n) \xrightarrow{P} V(\theta)$ par continuité. Ainsi $V^{-1/2}(T_n) \xrightarrow{P} V^{-1/2}(\theta)$ et on conclut en utilisant Slutsky que la loi limite de $\sqrt{n}\|V^{-1/2}(T_n)(T_n-\theta)\|^2$ est une χ^2_d. □

Remarque 26 *La seconde fonction pivotale asymptotique $\sqrt{n}\|V^{-1/2}(T_n)(T_n-\theta)\|^2$ est toujours inversible en θ. Elle est plus souvent utile que $\sqrt{n}\|V^{-1/2}(\theta)(T_n-\theta)\|^2$ qui n'est valable que lorsque la fonction V de Θ dans l'ensemble des matrices $d\times d$ symétrique définie positive est inversible (en tant que fonction et non en tant que matrice).*

Définition 7.4.1 *La fonction pivotale asymptotique de Wald vaut*

$$P_n^W(\theta) = n(\hat{\theta}_n-\theta)^T I(\hat{\theta}_n)(\hat{\theta}_n-\theta) = (\hat{\theta}_n-\theta)^T I_n(\hat{\theta}_n)(\hat{\theta}_n-\theta).$$

Une application du théorème 7.4.1 fournit la loi limite de $Q_n^W(\theta)$:

Corollaire 7.4.1 *Si le modèle (P_θ,Θ) est régulier et identifiable et la REV $\hat{\theta}_n$ unique à partir d'un certain rang alors la loi limite de $Q_n^W(\theta)$ est une χ^2_d où d est la dimension de θ, i.e. $\Theta\subseteq\mathbb{R}^d$.*

Démonstration : On applique le théorème 7.4.1 en remarquant que les hypothèses de régularité du modèle assure que l'information de Fisher soit une fonction continue. □

A partir de cette fonction pivotale asymptotique, il est facile de construire des régions de confiance de taille asymptotique $1-\alpha$.

Exemple 7.4.1 *Soit le modèle Gaussien $(\mathcal{N}(\mu,\sigma^2), \theta=(\mu,\sigma^2)\in\mathbb{R}\times\mathbb{R}_+^*)$ alors l'unique REV vaut $\hat{\theta}_n = (\overline{X}_n, S_n^2)$ et l'information de Fisher vaut*

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

La fonction pivotale asymptotique de Wald vaut donc

$$Q_n^W(\theta) = n\left(\frac{(\overline{X}_n-\mu)^2}{S_n^2} + \frac{(S_n^2-\sigma^2)^2}{2(S_n^2)^2}\right)$$

Ainsi, la région de confiance de taille asymptotique $1-\alpha$ est fournie par

$$\left\{(x,y)\in\mathbb{R}^2 \ / \ \frac{(\overline{X}_n-x)^2}{S_n^2} + \frac{(S_n^2-y)^2}{2(S_n^2)^2} \le \frac{q_{1-\alpha}^{\chi^2_2}}{n}\right\}.$$

C'est l'intérieur d'une ellipse dans le plan, centrée en $(\overline{X}_n, S_n^2)$.

Troisième partie

Tests d'hypothèses

Chapitre 8

Introduction aux tests paramétriques

8.1 Problématique de test

Soit un modèle paramétrique (P_θ, Θ) où le paramètre θ est inconnu. Le statisticien ne cherche pas directement à inférer la valeur θ mais plutôt de savoir si θ appartient à un ensemble de paramètres $\Theta_0 \subsetneq \Theta$: l'objectif d'un test est de décider si $\theta \in \Theta_0$, ou pas.

Exemple 8.1.1 *Une des premières applications de la théorie des tests était liée au problème militaire de détection de la présence d'un missile à l'aide d'un radar. L'écho d'un radar est "grand" si un missile est présent et il est "petit" dans le cas contraire. Supposons qu'on observe un échantillon $(X_1, \ldots, X_n)$ d'échos de radar aux instants successifs $1, \ldots, n$. Le caractère aléatoire de ces échos est lié aux effets de bruit de propagation d'ondes, des erreurs de mesure, etc... On se place dans le cadre d'un modèle paramétrique où $(X_1, \ldots, X_n)$ est issu d'un modèle P_θ avec θ inconnu et soit Θ_0 l'ensemble des paramètres correspondant à un écho suffisamment "grand". Le problème est alors de décider à partir de l'échantillon si oui ou non $\theta \in \Theta_0$, i.e. si oui ou non un missile est présent.*

8.1.1 Premières définitions

Soit $\Theta_0 \subsetneq \Theta$ et $\Theta_1 = \Theta \setminus \Theta_0$ (alors Θ_0 et Θ_1 forment une partition de Θ). On utilise l'écriture symbolique suivante pour définir le problème de test

$$H_0 : \quad \theta \in \Theta_0 \qquad H_1 : \quad \theta \in \Theta_1$$

où H_0 est l'hypothèse nulle et H_1 l'hypothèse alternative. Chacune de ces hypothèses peut être de deux natures :

Définition 8.1.1 *Pour $i = 0$ ou $= 1$ si H_i correspond à un ensemble Θ_i réduit à un singleton $\{\theta_i\}$ alors l'hypothèse H_i est dite simple. Dans le cas contraire, l'hypothèse est composite.*

Etant donné l'hypothèse nulle $H_0 : \quad \theta \in \Theta_0$ construire une procédure de test revient à construire à partir de l'échantillon $(X_1, \ldots, X_n)$ une règle de décision ϕ_n qui indique si oui ou non H_0 est vérifiée. Formellement, on a la définition

Définition 8.1.2 *Un test simple est une fonction mesurable $\phi_n : \mathcal{X}^n \to \{0, 1\}$ qui ne dépend pas de θ. On accepte l'hypothèse nulle H_0 lorsque $\phi_n = 0$ et on la rejette lorsque $\phi_n = 1$, i.e. on accepte l'hypothèse alternative H_1.*

Un test randomisé est une fonction mesurable $\phi_n : \mathcal{X}^n \to [0, 1]$ qui ne dépend pas de θ. Lorsque $\phi_n \in \{0, 1\}$, les règles de décision sont les mêmes que pour les tests (si $\phi_n = 0$ on accepte l'hypothèse nulle, si $\phi_n = 1$ on la rejette). Lorsque $\phi_n \in]0, 1[$, on rejette l'hypothèse nulle avec la probabilité ϕ_n et on l'accepte donc avec probabilité $1 - \phi_n$.

Un test simple ϕ_n est une v.a. ne prenant que 2 valeurs, 0 ou 1, c'est donc une variable de Bernoulli. On appelle zone de rejet du test l'ensemble $R_n = \{\phi_n((X_1, \ldots, X_n)) = 1\}$, i.e. la zone des observations qui conduisent à rejeter l'hypothèse nulle. Evidemment, construire un test simple est équivalent à donner une zone de rejet R_n car alors le test s'écrit de manière unique $\phi_n(X_1, \ldots, X_n) = 1_{R_n}(X_1, \ldots, X_n)$.

Remarque 27 *Par définition d'une statistique exhaustive T_n, elle contient toute l'information de l'échantillon pour inférer θ. On recherche donc une zone de rejet R sous la forme $R = \{T_n \in C_n\}$ pour un ensemble C_n à déterminer.*

8.1.2 Risques des tests

Ayant construit un test, on prend la décision d'accepter ou non H_0 à partir de l'échantillon observé $(X_1, \ldots, X_n)$. Il y a 4 possibilités :

- On accepte à raison H_0, i.e. $\phi_n(X_1, \ldots, X_n) = 0$ et $\theta \in \Theta_0$,
- On rejette à raison H_0, i.e. $\phi_n(X_1, \ldots, X_n) = 1$ et $\theta \in \Theta_1$,
- On rejette à tort H_0, i.e. $\phi_n(X_1, \ldots, X_n) = 1$ et $\theta \in \Theta_0$,
- On accepte à tort H_0, i.e. $\phi_n(X_1, \ldots, X_n) = 0$ et $\theta \in \Theta_1$.

On parlera dans les 2 derniers cas d'erreurs de tests lié au fait qu'on prend une décision sur le paramètre θ inconnu à partir des observations $(X_1, \ldots, X_n)$ uniquement. Rejeter à tort H_0 correspond à l'erreur de premier espèce et accepter à tort H_0 l'erreur de second espèce.

Le but du statisticien est de construire un test qui conduit à une erreur dans le moins de cas possibles.

Définition 8.1.3
- *Le risque de première espèce d'un test* ϕ_n *vaut* $\sup_{\theta\in\Theta_0} \mathbb{E}_\theta(\phi_n)$.
- *La fonction puissance d'un test est la fonction* $\pi : \Theta \to [0,1]$ *définie par la relation* $\pi(\theta) = \mathbb{E}_\theta(\phi_n)$ *pour tout* $\theta \in \Theta_1$.
- *Le risque de seconde espèce d'un test* ϕ_n *vaut* $\sup_{\theta\in\Theta_1} 1 - \pi(\theta)$.
- *La puissance d'un test* ϕ_n *est la fonction* $\pi(\theta)$ *restreinte à l'ensemble* Θ_1.

Remarque 28 *Dans le cas d'un test simple* $\phi_n = 1_{R_n}$ *alors le risque de première espèce est la plus grande probabilité de rejeter à tort (commettre une erreur de première espèce), i.e.* $\sup_{\theta\in\Theta_0} P_\theta(R_n) = \sup_{H_0} P_\theta(\,''On\ rejette\ H_0\,'')$. *Le risque de seconde espèce est la plus grande probabilité d'accepter à tort (commettre une erreur de seconde espèce), i.e.* $\sup_{\theta\in\Theta_1} 1 - P_\theta(R_n) = \sup_{H_1} P_\theta(\,''On\ accepte\ H_0\,'')$.

Le but du statisticien est donc de construire un test dont les risques de première et seconde espèce sont les plus faibles possibles (ou de manière équivalente un test dont le risque de première espèce est faible et la puissance est forte).

8.1.3 Approche de Neyman et niveau d'un test

Réduire le risque de première espèce conduit malheureusement souvent à augmenter le risque de seconde espèce. Ainsi, le test $\phi_n = 1_\emptyset$ qui accepte toujours H_0 ne commet jamais d'erreur de première espèce car il ne rejette jamais. Par contre, sa fonction puissance π est nulle sur Θ_1 et donc son risque de seconde espèce vaut 1 : quand l'hypothèse alternative H_1 est satisfaite, on commet une erreur de seconde espèce systématique en acceptant H_0.

Le principe de Neyman est de se fixer un seuil de tolérance sur le risque de première espèce appelé niveau :

Définition 8.1.4 *On dit qu'un test* ϕ_n *est de niveau* $\alpha \in [0,1]$ *si son risque de première espèce est inférieur ou égal à* α, *i.e.* $\sup_{\theta\in\Theta_0} \mathbb{E}_\theta(\phi_n) \leq \alpha$. *Le test* ϕ_n *est de taille* α *si son risque de première espèce est égal à* α, *i.e.* $\sup_{\theta\in\Theta_0} \mathbb{E}_\theta(\phi_n) = \alpha$. *On note alors* $R_n(\alpha)$ *la zone de rejet du test simple* ϕ_n *de taille* α :

$$\phi_n = 1_{R_n(\alpha)} \Leftrightarrow \sup_{\theta\in\Theta_0} \mathbb{P}_\theta(R_n(\alpha)) = \alpha$$

Parmi les tests d'un niveau α fixé il faut ensuite choisir celui qui a la plus grande puissance π, i.e. le plus petit risque de second espèce.

Définition 8.1.5 *Soit* $\alpha \in [0,1]$ *et doit un test* ϕ_n *de niveau* α. *Le test* ϕ_n *est sans biais si* $\pi(\theta) \geq \alpha$ *pour tout* $\theta \in \Theta_1$. *Il est uniformément plus puissant (UPP) si pour tout test* ϕ'_n *de niveau* α *et de puissance* π' *on a* $\pi(\theta) \geq \pi'(\theta)$ *pour tout* $\theta \in \Theta_1$.

Le principe de Neyman est de trouver un test UPP pour un niveau α qui est fixé par le statisticien.

Exemple 8.1.2 *On reprend l'exemple des missiles où on suppose que l'écho d'un radar suit le modèle Gaussien* $(\mathcal{N}(\theta,1), \theta \in \mathbb{R})$. *On veut tester si il y a un missile ou non soit*

$$H_0 : \quad "Il\ y\ a\ un\ missile" : \theta \geq \theta^* \qquad H_1 : \quad "Il\ n'y\ a\ pas\ de\ missile" : \theta < \theta^*$$

où θ^* *est connu avec un niveau de* 5%. *Les deux hypothèses sont composites. On sait que* $\overline{X}_n$ *est une statistique exhaustive pour* θ *(c'est aussi l'EMV sans biais de variance minimale). On construit un test simple dont la zone de rejet est* $R_n = \{\overline{X}_n < C\}$ *où* C *est une constante à déterminer. Comme* $\overline{X}_n \sim \mathcal{N}(\theta, n^{-1})$, *on calcule :*

$$P_\theta(R) = \mathbb{P}_\theta(\overline{X}_n < C) = \mathbb{P}_\theta(\theta + N/\sqrt{n} < C) = \Phi(\sqrt{n}(C-\theta))$$

où $N \sim \mathcal{N}(0,1)$ *et* Φ *est la fonction de répartition associée. Pour que le test* $\phi_n = 1_{\{\overline{X}_n < C\}}$ *est un niveau* $\alpha = 0.05$, *il faut donc la relation*

$$\sup_{\theta \in \Theta_0} \Phi(\sqrt{n}(C-\theta)) \leq 0.05.$$

Comme toute fonction de répartition, Φ *est croissante donc de manière équivalente*

$$\sup_{\theta > \theta^*} \sqrt{n}(C-\theta) \leq q^N_{0.05} \Leftrightarrow \sqrt{n}(C-\theta^*) \leq -1.64 \Leftrightarrow C \leq \theta^* - 1.64/\sqrt{n}.$$

Parmi tous les tests $\phi_n = 1_{\{\overline{X}_n < C\}}$ *de niveau* 0.05 *(qui vérifient* $C \leq \theta^* - 1.64/\sqrt{n}$*) on va choisir celui qui est le plus puissant. On calcule la fonction puissance* $\pi(\theta) = P_\theta(R) = \Phi(\sqrt{n}(C-\theta))$ *qui est croissante avec* C. *Donc le test qui a la plus grande puissance parmi les tests de la forme* $\phi_n = 1_{\{\overline{X}_n < C\}}$ *est celui qui est associé à la plus grande valeur de* C *qui assure un niveau* 0.05 *soit*

$$\phi_n = 1_{\{\overline{X}_n < \theta^* - 1.64/\sqrt{n}\}}.$$

8.1.4 p-valeur

En pratique, accepter ou rejeter l'hypothèse nulle n'a que peu de valeur scientifique : il suffit de baisser la valeur du niveau α jusqu'à accepter le test (le seul test de risque de premier espèce égal à 0 est le test $\phi = 0$ qui accepte toujours l'hypothèse nulle !). D'où la définition suivante

Définition 8.1.6 *La p-valeur d'une famille de tests de zones de rejet $R_n(\alpha)$, $0 < \alpha < 1$, est le plus petit niveau α^* pour lequel on rejette H_0, i.e si $(x_1, \ldots, x_n)$ est une réalisation de $(X_1, \ldots X_n)$ alors la p-valeur vaut $\alpha^* = \inf\{\alpha \in]0,1[\ / \ (x_1, \ldots, x_n) \in R_n(\alpha)\}$.*

La p-valeur (p-value en anglais) est fournie en sortie des procédures de tests dans le logiciel R.

Remarque 29

1. *Si la p-valeur est plus petite que 1%, on rejette H_0 pour tous les niveaux de tests "classiques" (en général α est choisi parmi 1, 5 ou 10%). Si la p-valeur est comprise entre 1% et 5% on a tendance à rejeter H_0, si elle est entre 5% et 10%, on rejette H_0 prudemment. Dans tous ces cas, on dit que le test est significatif car il permet de prendre une décision (rejeter H_0) avec une grande probabilité que H_1 soit vérifiée.*
2. *On a tendance à accepter H_0 si la p-valeur est supérieure à 10%. Mais alors H_0 n'est pas forcément vérifiée avec grande probabilité : il peut y avoir des "faux positifs" c'est à dire des cas où on accepte H_0 alors que H_1 est vérifiée. La p-valeur (risque de première espèce) ne suffit pas pour prendre une décision vraie avec grande probabilité lorsque celle-ci est grande (supérieur à 10%). On dit alors qu'on accepte H_0 mais que le test n'est pas significatif. Le calcul du risque de second espèce (ou de la puissance) nous permet de quantifier cette erreur de second espèce.*

8.1.5 Dualité entre régions de confiance et tests

Il existe une dualité entre régions de confiance et tests. Elle permet de construire facilement des tests d'un niveau donné à partir des régions de confiance classiques. Par contre elle ne donne aucun renseignement sur la puissance du test (ni son risque de seconde espèce).

On rappelle qu'une région de confiance C_n de niveau $1-\alpha$ est définie par la relation

$$P_\theta(\theta \in C_n) \geq 1 - \alpha, \qquad \forall \theta \in \Theta.$$

Soit alors le problème de test hypothèse simple-hypothèse composite de la forme

$$H_0 : \quad \theta = \theta_0 \qquad H_1 : \quad \theta \neq \theta_0$$

pour $\theta_0 \in \Theta$ connu.

Proposition 8.1.1 *Le test simple ϕ_n associé à la zone de rejet $R_n = \{\theta_0 \notin C_n\}$ est un test de niveau α.*

Démonstration : Il suffit de calculer le risque de première espèce

$$\sup_{\theta\in\Theta_0} P_\theta(\theta_0 \notin C_n) = P_{\theta_0}(\theta_0 \notin C_n) = 1 - P_{\theta_0}(\theta_0 \in C_n) \leq \alpha$$

par définition de la région de confiance. □

Exemple 8.1.3 *Soit le modèle Gaussien* $(\mathcal{N}(\theta, \sigma^2), \theta \in \mathbb{R})$ *avec* $\sigma^2 > 0$ *connu, alors on a l'intervalle de confiance de taille* $1 - \alpha$

$$C_n = \left[\overline{X}_n - \frac{\sigma}{\sqrt{n}} q^N_{1-\alpha/2}, \overline{X}_n + \frac{\sigma}{\sqrt{n}} q^N_{1-\alpha/2}\right].$$

On en déduit immédiatement un test ϕ_n *de niveau* α *pour le problème hypothèse simple-hypothèse composite*

$$H_0 : \quad \theta = \theta_0 \qquad H_1 : \quad \theta \neq \theta_0$$

de zone de rejet

$$\left\{|\overline{X}_n - \theta_0| > \frac{\sigma}{\sqrt{n}} q^N_{1-\alpha/2}\right\}.$$

On peut généraliser le problème de test à celui de $(\theta_i)_{i\in I} = (\theta_{0,i})_{i\in I}$ contre $(\theta_i)_{i\in I} \neq (\theta_{0,i})_{i\in I}$ où $(\theta_i)_{i\in I}$ est un ensemble de coordonnées de θ. Pour cela on utilise une RC pour $(\theta_i)_{i\in I}$ avec les paramètres $(\theta_i)_{i\notin I}$ inconnus.

Exemple 8.1.4 *Soit le modèle Gaussien* $(\mathcal{N}(\theta, \sigma^2), \theta = (\mu, \sigma^2) \in \mathbb{R}) \times \mathbb{R}^*_+$ *avec le problème de test portant uniquement sur* $\mu = \theta_1$ *de la forme*

$$H_0 : \quad \mu = \mu_0 \qquad et \qquad H_1 : \quad \mu \neq \mu_0.$$

Remarquons qu'on peut réécrire le problème de test sous la forme

$$H_0 : \quad \theta = \mu_0 \times \mathbb{R}^*_+ \qquad et \qquad H_1 : \quad \theta \neq \mu_0 \times \mathbb{R}^*_+.$$

C'est donc un test hypothèse nulle composite contre hypothèse alternative composite. Un test est construit à partir de l'IC de taille $1 - \alpha$ *pour* μ *avec* $\theta_2 = \sigma^2$ *inconnu de la forme*

$$C_n = \left[\overline{X}_n - \frac{\sqrt{S_n^{2\prime}}}{\sqrt{n}} q^{T_{n-1}}_{1-\alpha/2}, \overline{X}_n + \frac{\sqrt{S_n^{2\prime}}}{\sqrt{n}} q^{T_{n-1}}_{1-\alpha/2}\right].$$

On en déduit le test ϕ_n *de niveau* α *déterminé par la zone de rejet*

$$\left\{|\overline{X}_n - \theta_0| > \frac{\sqrt{S_n^{2\prime}}}{\sqrt{n}} q^{T_{n-1}}_{1-\alpha/2}\right\}.$$

On parle alors du test de Student.

8.2 Tests asymptotiques

Tout comme pour les régions de confiance, il n'est pas toujours possible de construire un test de taille α donné. On fait alors appel à l'asymptotique.

8.2.1 Niveau asymptotique

Définition 8.2.1 *Soit $\alpha \in [0,1]$, la suite de test (ϕ_n) est de niveau (taille) asymptotique α si*

$$\forall \theta \in \Theta_0, \quad \limsup_{n\to\infty} P_\theta(\phi_n = 1) \leq \ (=)\ \alpha.$$

La suite de test (ϕ_n) est convergente si sa puissance asymptotique vaut 1 :

$$\forall \theta \in \Theta_1, \quad \lim_{n\to\infty} P_\theta(\phi_n = 1) = 1.$$

La p-valeur asymptotique d'une famille de tests de zones de rejet $R_n(\alpha)$, $0 < \alpha < 1$, est le plus petit niveau asymptotique α^* pour lequel on rejette H_0.

Asymptotiquement, il n'y a pas de tests plus puissants qu'un test convergent.

Exemple 8.2.1 *Soit le modèle $(P_\theta, \theta = \mathbb{E}_\theta(X) \in \mathbb{R})$ avec le problème de test $H_0 : \theta = \theta_0$ contre $H_1 : \theta \neq \theta_0$.*

- *Si $Var_\theta(X) = \sigma^2$ est connu, alors on est dans un cas d'hypothèse nulle simple, on peut utiliser la dualité entre tests et régions de confiance. On a*

$$C_n = [\overline{X}_n - \frac{\sigma}{\sqrt{n}} q^N_{1-\alpha/2}, \overline{X}_n + \frac{\sigma}{\sqrt{n}} q^N_{1-\alpha/2}]$$

qui est de taille asymptotique $1-\alpha$ donc la suite des tests ϕ_n de zone de rejet

$$R_n = \{|\sqrt{n}\overline{X}_n - \theta_0| > \sigma q^N_{1-\alpha/2}\}$$

est de niveau asymptotique α. La puissance de ce test converge vers 1 pour tout $\theta \in \theta_1$:

$$P_\theta\left(\sqrt{n}|\overline{X}_n - \theta_0| > \sigma q^N_{1-\alpha/2}\right) \to 1,$$

car la LFGN implique que $\overline{X}_n - \theta_0 \xrightarrow{P} \theta - \theta_0 \neq 0$. Donc la suite de tests est convergente.

- *Si $Var_\theta(X) = \sigma^2$ est inconnue alors on obtient une suite de tests avec les mêmes propriétés que précédemment en remplaçant simplement σ par $\sqrt{S^2_n}$, un estimateur consistant de la variance.*

Nous allons donner 2 exemples de tests asymptotiques convergents dans le problème

$$H_0 : \quad g(\theta) = 0_k \qquad H_1 : \quad g(\theta) \neq 0_k$$

où la fonction $g : \Theta \to \mathbb{R}^k$ avec $1 \leq k \leq d$ est connue est satisfait

(HG) La fonction g est continûment différentiable telle que sa Jacobienne $J_\theta g(\theta)$ soit de plein rang k.

On retrouve les tests à hypothèse nulle simple avec $k = d$ et $g(\theta) = \theta - \theta_0$.

8.2.2 Test de Wald

Soit un modèle régulier identifiable dans lequel il existe une unique REV $\hat{\theta}_n$.

Le test de Wald est construit à partir de l'efficacité asymptotique de $(\hat{\theta}_n)$. En effet, en appliquant la δ-méthode à g, on trouve

$$\sqrt{n}(g(\hat{\theta}_n) - g(\theta)) \xrightarrow{\mathcal{L}} \mathcal{N}_k(0_k, J_\theta g(\theta) I^{-1}(\theta) J_\theta g(\theta)^T).$$

On a donc un estimateur asymptotiquement normal de $g(\theta)$ et en utilisant le théorème 7.4.1 on trouve que

$$n(g(\hat{\theta}_n) - g(\theta))^T (J_\theta g(\hat{\theta}_n) I^{-1}(\hat{\theta}_n) J_\theta g(\hat{\theta}_n)^T)^{-1} (g(\hat{\theta}_n) - g(\theta)) \xrightarrow{\mathcal{L}} \chi^2_k.$$

Sous H_0, comme $g(\theta) = 0$ on en déduit alors que

$$\zeta_n^W := n g(\hat{\theta}_n)^T (J_\theta g(\hat{\theta}_n) I^{-1}(\hat{\theta}_n) J_\theta g(\hat{\theta}_n)^T)^{-1} g(\hat{\theta}_n) \xrightarrow{\mathcal{L}} \chi^2_k,$$

où ζ_n^W est appelé la statistique de Wald. Ce résultat asymptotique nous permet de construire une suite de tests de niveau asymptotique α :

Théorème 8.2.1 *Soit un modèle régulier identifiable dans lequel il existe une unique REV $\hat{\theta}_n$, alors la suite de tests (ϕ_n) de zone de rejet*

$$R_n = \{\zeta_n^W > q_{1-\alpha}^{\chi^2_k}\}$$

est de niveau asymptotique α et convergente.

Démonstration : Par construction la suite (ϕ_n) est de niveau asymptotique $1 - \alpha$. Nous montrons que sous H_1 on a $\zeta_n^W \xrightarrow{P} +\infty$ ce qui implique que $\lim P_\theta(\zeta_n^W > q_{1-\alpha}^{\chi^2_k}) = 1 \ \forall \theta \in \Theta_1$ et donc que le test et convergent.

Soit $V(\theta) = J_\theta g(\theta) I^{-1}(\theta) J_\theta g(\theta)^T$ la variance asymptotique de $g(\hat{\theta}_n)$. Alors on peut écrire

$$\zeta_n^W = n g(\hat{\theta}_n)^T V^{-1}(\hat{\theta}_n) g(\hat{\theta}_n) = T_{1,n} + T_{2,n} + T_{3,n}$$

avec

$$T_{1,n} = n g(\theta)^T V^{-1}(\hat{\theta}_n) g(\theta), \quad T_{2,n} = n(g(\hat{\theta}_n) - g(\theta))^T V^{-1}(\hat{\theta}_n)(g(\hat{\theta}_n) - g(\theta)),$$
$$T_{3,n} = 2n(g(\hat{\theta}_n) - g(\theta))^T V^{-1}(\hat{\theta}_n) g(\theta).$$

Comme $g(\hat{\theta}_n)$ est asymptotiquement normal pour $g(\theta)$, il est aussi fortement convergent donc $T_{2,n}/T_{1,n} = g(\hat{\theta}_n) - g(\theta))^T M(\hat{\theta}_n) g(\hat{\theta}_n) - g(\theta)) \xrightarrow{p.s.} 0$ car la matrice $k \times k$ $M(\hat{\theta}_n) \in \mathbb{R}^k$ converge p.s. par continuité vers $M(\theta) < \infty$. De même, $T_{3,n}/T_{1,n} = (g(\hat{\theta}_n) - g(\theta))^T K(\hat{\theta}_n) \xrightarrow{p.s.} 0$ car le vecteur $K(\hat{\theta}_n) \in \mathbb{R}^k$ converge p.s. par continuité vers $K(\theta) < \infty$. En réécrivant $\zeta_n^W = T_{1,n}(1 + T_{2,n}/T_{1,n} + T_{3,n}/T_{1,n})$ et comme $g(\theta)^T V^{-1}(\hat{\theta}_n) g(\theta) \xrightarrow{p.s.} g(\theta)^T V^{-1}(\theta) g(\theta) > 0$ car $g(\theta) \neq 0$ sous $H1$ et $V^{-1}(\theta) > 0$, on a finalement que

$$\zeta_n^W = n g(\theta)^T V^{-1}(\hat{\theta}_n) g(\theta)(1 + T_{2,n}/T_{1,n} + T_{3,n}/T_{1,n}) \xrightarrow{p.s.} \infty. \square$$

8.2.3 Test du score

Le principe du test du score provient de la remarque suivante : pour construire une suite de tests d'un niveau asymptotique donné, le comportement asymptotique sous H_0 suffit. On va donc se placer dans le modèle contraint (P_θ, Θ_0) supposer qu'il existe un unique EMV contraint $\hat{\theta}_n^0$.

Remarque 30 *Si l'hypothèse nulle est simple $\Theta_0 = \{\theta_0\}$ alors nécessairement $\hat{\theta}_n^0 = \theta_0$. Si Θ_0 est un intervalle, $\hat{\theta}_n^0$ est l'unique solution du système suivant, appelé l'EV contrainte :*

$$\nabla_{\theta,\lambda} H_n(\hat{\theta}_n^0, \hat{\lambda}_n^0) = 0$$

issu de la dérivation du Lagrangien $H_n(\theta, \lambda) = l_n(\theta) - \lambda^T g(\theta)$ où $\lambda \in \mathbb{R}^k$ et $\hat{\theta}_n^0$ vérifie la condition du second ordre.

Comme $g(\hat{\theta}_n^0) = 0$ car $\hat{\theta}_n^0 \in \Theta_0$ on ne peut pas utiliser le même raisonnement que pour le test de Wald directement. On rappelle la définition du vecteur score de l'échantillon

$$S_n(\theta) = \nabla_\theta \log(L_n(\theta))$$

et on définit la statistique du score

$$\zeta_n^S := n^{-1} S_n(\hat{\theta}_n^0)^T I^{-1}(\hat{\theta}_n^0) S_n(\hat{\theta}_n^0).$$

Cette statistique ne nécessite que le calcul de l'EMV contraint $\hat{\theta}_n^0$ et on a

Théorème 8.2.2 *Soit un modèle régulier identifiable dans lequel il existe un unique EMV contraint $\hat{\theta}_n$, alors la suite de tests (ϕ_n) de zone de rejet*

$$R_n = \{\zeta_n^S > q_{1-\alpha}^{\chi_k^2}\}$$

est de niveau asymptotique α et convergente.

Démonstration : Pour tout $\theta \in \Theta$, la loi asymptotique du score vaut :

$$\frac{1}{\sqrt{n}} S_n(\theta) \xrightarrow{\mathcal{L}} \mathcal{N}_d\left(0, I(\theta)\right)$$

Donc $\frac{1}{\sqrt{n}} I(\theta)^{-1/2} S_n(\theta)$ converge vers un vecteur gaussien isotrope de $\mathbb{R}^d$. On montre que $\frac{1}{\sqrt{n}} I(\hat{\theta}_n^0)^{-1/2} S_n(\hat{\theta}_n^0)$ est la projection orthogonale de ce vecteur sur Θ_0 de dimension k. Donc en appliquant le Théorème de Cochran on obtient

$$n^{-1} S_n(\hat{\theta}_n^0) I(\hat{\theta}_n^0)^{-1} S_n(\hat{\theta}_n^0)^T \xrightarrow{\mathcal{L}} \chi_2^k.$$

Le niveau asymptotique de la zone de rejet en découle facilement.

Sous (H1), comme $\hat{\theta}_n$ est fortement convergent et que $\theta \notin \Theta_0$ on a $\liminf \|\hat{\theta}_n - \hat{\theta}_n^0\| \geq \varepsilon$ pour $\varepsilon > 0$. Comme par définition on a les relations $\nabla l_n(\hat{\theta}_n) = 0$ et $n \nabla l_n(\theta) = -S_n(\theta)$ on obtient

$$S_n(\hat{\theta}_n^0) = n \mathbb{H}_\theta(l_n(\hat{\theta}_n^0))(\hat{\theta}_n - \hat{\theta}_n^0) + o(\hat{\theta}_n - \hat{\theta}_n^0).$$

Le comportement asymptotique de ζ_n^S est donc le même que

$$n(\hat{\theta}_n - \hat{\theta}_n^0)^T \mathbb{H}_\theta(l_n(\hat{\theta}_n^0)) I(\hat{\theta}_n^0)^{-1} \mathbb{H}_\theta(l_n(\hat{\theta}_n^0))(\hat{\theta}_n - \hat{\theta}_n^0)$$

et $\zeta_n^S \sim n(\hat{\theta}_n - \hat{\theta}_n^0)^T I(\hat{\theta}_n^0)(\hat{\theta}_n - \hat{\theta}_n^0) \xrightarrow{p.s.} +\infty$ en appliquant la loi forte des grands nombres uniforme à $\theta \to S(X_i, \theta)$ car d'après (H4) : $I(\hat{\theta}_n^0) > 0$ pour tout $n \in \mathbb{N}$. On conclut que le test est convergent en suivant le même raisonnement que pour le test de Wald. □

Chapitre 9

Test du Rapport de Vraisemblance

9.1 Définition

Dans ce chapitre nous étudions le Test du Rapport de Vraisemblance (TRV) et ses propriétés (non-)asymptotiques pour la problématique de test $H_0 : \theta \in \Theta_0$ contre $H_1 : \theta \in \Theta_1$. Pour se faire, on suppose que le modèle paramétrique $(P_\theta, \Theta = \Theta_0 \cup \Theta_1)$ satisfait l'hypothèse usuelle (H1) satisfaite : le support de la loi ne dépend pas de θ.

Définition 9.1.1 *On appelle TRV tout test construit à l'aide du rapport de vraisemblance (RV) définit en tous points $a \in \Theta_0$ et $b \in \Theta_1$ par la relation*

$$V_{a,b} = \frac{L_n(b)}{L_n(a)} \quad si \quad L_n(a) \neq 0, \qquad V_{a,b} = 0 \quad sinon.$$

Remarque 31 *Le RV est bien défini car sous (H1) on a $L_n(a) = 0 \Rightarrow L_n(b) = 0$.*

9.2 Propriétés non asymptotiques

Nous avons vu comment obtenir des tests d'un niveau donné et des tests d'un niveau asymptotique donné convergent. Nous allons étudié la construction de tests d'un niveau donné UPP d'un niveau α donné. Ce problème complexe n'a pas toujours de solution, il faut spécifier dans quel problématique de test on se place.

9.2.1 Lemme de Neyman-Pearson

On se place dans un problème de test hypothèse nulle simple $\Theta_0 = \{\theta_0\}$ contre hypothèse alternative simple $\Theta_1 = \{\theta_1\}$ (par définition $\Theta = \{\theta_0\} \cup \{\theta_1\}$). Le RV s'écrit

$$V_{\theta_0,\theta_1} = L_n(\theta_1)/L_n(\theta_0) 1_{L_n(\theta_0)\neq 0}$$

Pour cette problématique de test on appelle test du rapport de vraisemblance (abrégé en TRV) tout test $\phi_{C,c}$ de la forme
- $\phi_{C,c} = 1$ si $L_n(\theta_1) > CL_n(\theta_0)$,
- $\phi_{C,c} = 0$ si $L_n(\theta_1) < CL_n(\theta_0)$ et
- $\phi_{C,c} = c \in]0,1[$ si $L_n(\theta_1) = CL_n(\theta_0)$.

pour $C > 0$ et $c \in [0,1]$ à fixer.

Remarque 32 *En pratique, on rejettera toujours H_0 lorsque θ_1 est plus vraisemblable que θ_0, i.e. $L_n(\theta_1) > L_n(\theta_0)$ donc on choisira $C \geq 1$.*

On a alors le résultat fondamental suivant

Lemme 9.2.1 (Neyman-Person) *Soit $\alpha \in]0,1[$ alors il existe des constantes C, c telles que le TRV $\phi_{C,c}$ soit de taille α et ce test est alors UPP de niveau α.*

Démonstration : On calcule le risque de premier espèce du TRV randomisé $\mathbb{E}_{\theta_0}(\phi_{C,c})$. On considère F la fonction de répartition de la variable aléatoire positive $L_n(\theta_1)/L_n(\theta_0)$ sous H_0 et on note C son quantile d'ordre $1-\alpha$. On distingue deux cas :
- Soit $P_{\theta_0}(L_n(\theta_1) = CL_n(\theta_0)) = 0$ et on considère le TRV simple $\phi_{C,0}$. On a directement $\mathbb{E}_{\theta_0}(\phi_{C,0}) = P_{\theta_0}(\phi_{C,0} = 1) = 1 - F(C) = \alpha$.
- Soit $P_{\theta_0}(L_n(\theta_1) = CL_n(\theta_0)) > 0$ et on considère le TRV randomisé $\phi_{C,c}$ avec c vérifiant
$$c = \frac{\alpha + F(C) - 1}{P_{\theta_0}(L_n(\theta_1) = CL_n(\theta_0))}.$$
On a bien $c > 0$ car $F(C) \geq 1 - \alpha$ par définition et

$$\begin{aligned}\mathbb{E}_{\theta_0}(\phi_{C,c}) &= P_{\theta_0}(\phi_{C,0} = 1) + cP_{\theta_0}(\phi_{C,0} = c) \\ &= 1 - F(C) + \frac{\alpha + F(C) - 1}{P_{\theta_0}(L_n(\theta_1) = CL_n(\theta_0))} P_{\theta_0}(L_n(\theta_1) = CL_n(\theta_0)) = \alpha\end{aligned}$$

On montre maintenant que $\phi_{C,c}$ est UPP. Soit ϕ un test de niveau α, i.e. tel que $\mathbb{E}_{\theta_0}(\phi) \leq \alpha$, on montre que $\mathbb{E}_{\theta_1}(\phi_{C,c} - \phi) \geq 0$, i.e. que la puissance de ϕ est plus faible que celle du TRV $\phi_{C,c}$. Notons que

$$\Delta = \mathbb{E}_{\theta_1}(\phi_{C,c} - \phi) - C\mathbb{E}_{\theta_0}(\phi_{C,c} - \phi) = (L_n(\theta_1)/L_n(\theta_0) - C)\mathbb{E}_{\theta_0}(\phi_{C,c} - \phi).$$

Si $\phi_{C,c}(x) = 0$ alors par définition $L_n(\theta_1)/L_n(\theta_0) - C < 0$ et $\phi_{C,c}(x) - \phi \leq 0$, et si $\phi_{k,c}(x) = 1$ alors $L_n(\theta_1)/L_n(\theta_0) - C > 0$ et $\phi_{C,c}(x) - \phi \geq 0$ car $\phi \in [0,1]$. Dans tous les cas $\Delta \geq 0$ et le résultat est prouvé. □

Remarque 33 *Si le modèle est absolument continu, i.e. admet une densité par rapport à la mesure de Lebesgue, alors le TRV simple $\phi_{C,0}$ de taille α est UPP.*

Exemple 9.2.1

– *Cas du modèle gaussien* $(\mathcal{N}(\theta,1), \theta \in \mathbb{R})$ *alors*

$$L_n(\theta) = (2\pi)^{-n/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(X_i-\theta)^2\right).$$

Pour simplifier on passe au logarithme (croissant) car (H1) est vérifiée et on obtient

$$\log \frac{L_n(\theta_1)}{L_n(\theta_0)} = (\theta_1-\theta_0)\sum_{i=1}^{n} X_i - \frac{n}{2}(\theta_1^2-\theta_0^2).$$

Si $\theta_1 > \theta_0$*, c'est une fonction croissante de* $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ *et on rejette l'hypothèse nulle* $\theta = \theta_0$ *dès que* $\overline{X}_n > C$*. On choisit alors* $C = \theta_0 + q^N_{1-\alpha}/\sqrt{n}$ *afin que le risque de premier espèce soit égal à* α *et un test UPP est donné par la zone de rejet*

$$\overline{X}_n > \theta_0 + q^N_{1-\alpha}/\sqrt{n}.$$

Si $\theta_0 > \theta_1$*, on trouve un test UPP avec la zone de rejet*

$$\overline{X}_n < \theta_0 - q^N_{1-\alpha}/\sqrt{n}.$$

– *Cas du modèle de Bernoulli* $(\mathcal{B}(\theta), 0<\theta<1)$*, si on pose* $T_n = X_1 + \cdots + X_n$ *la statistique exhaustive alors*

$$\frac{L_n(\theta_1)}{L_n(\theta_0)} = \left(\frac{1-\theta_1}{1-\theta_0}\right)^n \left(\frac{\theta_1}{\theta_0}\Big/\frac{1-\theta_1}{1-\theta_0}\right)^{T_n}.$$

C'est une fonction croissante de T_n *lorsque* $\theta_1 > \theta_0$*, donc la zone de rejet est de la forme* $T_n \geq k$*. Comme* $T_n \sim \mathcal{B}(n,\theta)$ *discrète à valeur entière, le TRV* $\phi_{C,c}$ *peut être randomisé avec* C *qui est le plus petit entier tel que* $\tilde{\alpha} := P_{\theta_0}(T_n \leq C) \geq 1-\alpha$ *et* $c = (\alpha - \tilde{\alpha})/P_{\theta_0}(T_n = C)$*. C'est le test UPP de niveau* α*.*

9.2.2 Rapport de vraisemblance monotone

Nous avons obtenu dans la section précédente des tests UPP, i.e. préférables à tous les tests de niveau α dans le cas hypothèse simple contre hypothèse simple. Ce cadre théorique n'a que peu de valeurs en pratique car il réduit l'ensemble des paramètres à $\Theta = \{\theta_0\} \cup \{\theta_1\}$. Nous allons nous placer dans des problèmes de tests hypothèse composites contre hypothèse composites de la forme $H_0 : \theta \leq \theta^*$ (ou $\theta \geq \theta^*$) contre $H_1 : \theta > \theta^*$ (ou $\theta < \theta^*$). On définit

Définition 9.2.1 *Le modèle est à rapport de vraisemblance monotone en une statistique $T_n \in \mathbb{R}$ (RVM en T_n) lorsque $V_{a,b}$ est une fonction croissante de T_n, i.e. il existe une fonction V croissante en sa première variable telle qu'on ait*

$$V_{a,b} = V(T_n, a, b) \qquad \forall a \in \Theta_0,\ b \in \Theta_1.$$

On définit le test randomisé $\phi_{C,c}$ (appelé aussi TRV) de la forme
- $\phi_{C,c} = 1$ lorsque $T_n > C$,
- $\phi_{C,c} = 0$ lorsque $T_n < C$,
- $\phi_{C,c}(x) = c$ lorsque $T_n = C$

pour $C \in \mathbb{R}$ et $c \in [0, 1]$.

Remarque 34 *Si T_n est une v.a. absolument continue alors $P_\theta(T_n = C) = 0$ et on considère les tests simples $\phi_{C,0} = 1_{T_n > c}$.*

Théorème 9.2.1 (Karlin-Rubin) *Soit un modèle paramétrique à RVM en T_n ((H1) satisfaite) et soit le niveau $\alpha \in]0, 1[$ alors il existe des constantes C, c telles que $\alpha = \sup_{\theta \in \Theta_0} \mathbb{E}_\theta(\phi_{C,c})$ et le TRV $\phi_{C,c}$ est UPP de niveau α.*

La preuve de ce théorème est de même nature que celle du lemme 9.2.1.

Exemple 9.2.2 *Reprenons l'exemple de la détection des missiles. Nous avions vu que le test*

$$\phi_n = 1_{\{\overline{X}_n < \theta^* - 1.64/\sqrt{n}\}}$$

était plus puissant que les tests de la forme $\phi_n = 1_{\{\overline{X}_n \leq C\}}$. Il est en fait UPP de niveau α car le modèle est à RVM en $T_n = -\overline{X}_n$ car

$$V_{a,b} = \exp((b-a)n\overline{X}_n - n/2(b^2 - a^2)), \qquad b - a < 0.$$

Le test ϕ_n est de la forme d'un TRV $\phi_{C,c}$ et son risque de première espèce $\sup_{\theta \in \Theta_0} \mathbb{E}_\theta(\phi_n)$ est égal à α.

Voici un autre exemple de problème de test où un test UPP existe. Soit le problème de la forme $\Theta_0 = \{\theta \in \Theta |\, \theta \leq \theta_1 \text{ ou } \theta \geq \theta_2\}$ pour $\theta_1 < \theta_2$ et $\Theta_1 =]\theta_1, \theta_2[$ On définit alors

Exemple 9.2.3 *On définit le modèle exponentiel généralisé $(P_\theta, \Theta \subset \mathbb{R})$ de la forme*

$$f(x, \theta) = c(\theta)h(x) \exp\left(\alpha(\theta)T(x)\right)$$

pour des fonctions $h \geq 0$, $c \in \mathbb{R}$, $T \in \mathbb{R}$ non constant et $\theta \mapsto \alpha(\theta)$ une fonction strictement croissante.

Théorème 9.2.2 (Lehmann) *Dans le modèle exponentiel généralisé, si on note* $T_n = \sum_{i=1}^n T(X_i)$ *alors un test UPP de niveau* α *est défini par* $\phi = 1$ *pour* $T_n \in]t_1, t_2[$, $\phi = 0$ *pour* $T_n \notin [t_1, t_2]$, *et* $\phi = c_i$ *pour* $T_n = t_i$ *lorsque* $i = 1, 2$. *Les constantes* t_i, c_i *vérifient les relations* $\mathbb{E}_{\theta_i}(\phi) = \alpha$ *pour* $i = 1, 2$.

Exemple 9.2.4 *Soit le cas Gaussien* $(\mathcal{N}(\theta, 1), \mathbb{R})$ *avec* $\Theta_0 = \{\theta \in \Theta |\, \theta \leq \theta_1$ *ou* $\theta \geq \theta_2\}$ *pour* $\theta_1 < \theta_2$ *et* $\Theta_1 =]\theta_1, \theta_2[$. *Alors le test de zone de rejet* $\{|\overline{X} - (\theta_1 + \theta_2)/2| < (\theta_2 - \theta_1)/2 + \varphi_{1-\alpha/2}/\sqrt{n}\}$ *est UPP de niveau* α.

Il n'existe pas d'autre exemple de problème de test pour lesquels un test UPP d'un niveau donné existe.

Exemple 9.2.5 *Il existe un test de même nature défini par* $\phi = 1$ *pour* $T_n \notin [t_1, t_2]$, $\phi = 0$ *pour* $T_n \in]t_1, t_2[$, *et* $\phi = c_i$ *pour* $T_n = t_i$ *lorsque* $i = 1, 2$ *(et de niveau exactement* α*) pour le test bilatère d'hypothèse* $H_0 : \theta \in [\theta_1, \theta_2]$ *contre* $H_1 : \theta \notin [\theta_1, \theta_2]$ *(en particulier* $H_0 : \theta = \theta^*$ *contre* $H_1 : \theta \neq \theta^*$*).*
Attention, *ce test n'est pas UPP de niveau* α. *Il n'est plus puissant que parmi les tests sans biais de niveau* α.

9.3 TRV : cas général

On se place désormais dans un modèle régulier (P_θ, Θ) identifiable et dans la problématique de test général $H_0 : \theta \in \Theta_0$ contre $H_1 : \theta \in \Theta_1$. Le but de cette section est de construire un test du rapport de vraisemblance dans ce contexte général. On pose

$$V = \frac{\sup_{\theta \in \Theta_1} L_n(\theta)}{\sup_{\theta \in \Theta_0} L_n(\theta)}. \tag{9.1}$$

Définition 9.3.1 *Le TRV simple consiste à rejeter l'hypothèse nulle* H_0 *pour des grandes valeurs de* V, *i.e.* $R_n = \{V > C\}$.

Ce test coïncide avec le TRV $\phi_{C,0}$ dans les cas d'hypothèses simple, de rapports de vraisemblance monotones ou de modèles exponentiels généralisés. On sait donc qu'il est UPP dans de nombreux cas.

Un tel test est difficile à mettre en place en général car la loi de V est inconnue. On considère plutôt :

Proposition 9.3.1 *On suppose qu'il existe un unique EMV* $\hat{\theta}_n$ *pour le modèle* (P_θ, Θ) *et un unique EMV contraint* $\hat{\theta}_n^0$ *pour le modèle contraint* (P_θ, Θ_0). *Le TRV simple* $\phi_{C,n}$ *dans le problème* $H_0 : \theta \in \Theta_0$ *contre* $H_1 : \theta \in \Theta_1$ *a pour zone de rejet*

$$R_n : \zeta_n^{RV} > C, \quad avec \;\; \zeta_n^{RV} := 2n(l_n(\hat{\theta}_n^0) - l_n(\hat{\theta}_n)) \;\; et \;\; C > 0.$$

Démonstration : On remarque que le test $\phi = 1_{V>C}$ de taille α est équivalent à $\phi' = 1_{V'>C'}$ de taille α avec

$$V' = \frac{\sup_{\theta\in\Theta} L_n(\theta)}{\sup_{\theta\in\Theta_0} L_n(\theta)} = \frac{L_n(\hat{\theta}_n)}{L_n(\hat{\theta}_n^0)}.$$

Comme $\zeta_n^{RV} = 2\log V' = 2n(l_n(\hat{\theta}_n^0) - l_n(\hat{\theta}_n))$ avec $x \to 2\log(x)$ croissante le fait de rejeter H_0 lorsque V' est trop grand revient à rejeter H_0 lorsque $2\log(V')$ est trop grand et on obtient l'équivalence de ϕ' et de $\phi_{C,n}$. □

On sait donc que le TRV simple $\phi_{n,C}$ est UPP dans les problèmes à hypothèses simples, de RVM ou de modèles exponentiels généralisés. De plus il est optimal asymptotiquement pour le problème de test $H_0 : g(\theta) = 0$ contre $H_1 : g(\theta) \neq 0$ avec $g : \Theta \mapsto \mathbb{R}^k$ qui satisfait l'hypothèse (HG) (donc $k \geq d$). Plus précisément on a le résultat suivant :

Proposition 9.3.2 *Si on choisit $C = q_{1-\alpha}^{\chi_k^2}$ alors la suite des TRV $(\phi_{C,n})$ est de niveau asymptotique α et convergente. On appelle ces tests les TRV asymptotiques.*

Démonstration : On rappelle la normalité asymptotique du vecteur score obtenu sous H_0 équivaut à

$$n(\hat{\theta}_n - \hat{\theta}_n^0)I(\hat{\theta}_n^0)(\hat{\theta}_n - \hat{\theta}_n^0) \xrightarrow{\mathcal{L}} \chi_k^2.$$

On effectue alors un développement de Taylor d'ordre 2 de $l_n(\hat{\theta}_n^0)$ au point $\hat{\theta}_n$ et on obtient, en remarquant que $\nabla_\theta l_n(\hat{\theta}_n) = 0$

$$\zeta_n^{RV} = 2n(l_n(\hat{\theta}_n^0) - l_n(\hat{\theta}_n)) = 2n(\hat{\theta}_n^0 - \hat{\theta}_n)^T \frac{\mathbb{H}_\theta l_n(\tilde{\theta}_n)}{2}(\hat{\theta}_n^0 - \hat{\theta}_n)$$

où $\tilde{\theta}_n$ est un point entre $\hat{\theta}_n^0$ et $\hat{\theta}_n$. Par la LFGN (uniforme) on sait que $\|\mathbb{H}_\theta l_n - I\|_K \xrightarrow{p.s.} 0$ où $\|\cdot\|_K$ est la norme uniforme sur un compact au voisinage de θ, comme $\hat{\theta}_n^0$ et $\hat{\theta}_n$ sont tous les 2 fortement convergents vers θ, ils appartiennent à K à partir d'un certain rang, d'où $\tilde{\theta}_n$ appartient aussi à K pour n suffisamment grand et $\mathbb{H}_\theta l_n(\tilde{\theta}_n) - I(\hat{\theta}_n^0) \xrightarrow{p.s.} 0$. On conclut par le théorème de Slutsky que $\zeta_n^{RV} \xrightarrow{\mathcal{L}} \chi_k^2$ et ainsi que la suite $(\phi_{C',n})$ est bien de niveau asymptotique α.

On montre que la suite est convergente de la même manière que pour le test du score. □

Il n'est pas toujours facile de déterminer la loi de ζ_n^{RV} et on peut alors faire appel à l'asymptotique. Une autre méthode possible est de trouver une statistique T_n plus simple telle que $\zeta_n^{RV} > C \Leftrightarrow T_n > C'$, i.e. telle que $\zeta_n^{RV} = \phi(T_n)$ avec ϕ strictement croissante puis de raisonner directement sur la loi de T_n.

Exemple 9.3.1 *On considère le modèle gaussien* $(\mathcal{N}(\mu,\sigma^2),(\mu,\sigma^2)\in\mathbb{R}\times\mathbb{R}_+^*)$ *et on cherche à tester l'hypothèse nulle composite* $\mu=\mu_0$ *contre l'hypothèse alternative composite* $\mu\neq\mu_0$ *avec* μ_0 *connu. On sait que*

$$L_n(\theta)=\frac{1}{(\sigma\sqrt{2\pi})^n}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n(X_i-\mu)^2\right)$$

admet une unique REV $\hat{\theta}_n=(\overline{X}_n,S_n^2)$ *qui vérifie les conditions du second ordre. Comme* Θ *est un intervalle ouvert, c'est l'unique EMV.*
Le modèle contraint $(\mathcal{N}(\mu,\sigma^2),\Theta=\{\mu_0\}\times\mathbb{R}_+^*)$ *admet lui aussi un unique EMV* $\hat{\theta}_n^0=(\mu_0,\overline{(X-\mu_0)^2}_n)$ *car*

$$\frac{\partial}{\partial\sigma^2}\sum_{i=1}^n\log f(X_i,\theta)=\frac{1}{2}\left(\frac{1}{\sigma^4}\sum_{i=1}^n(X_i-\mu_0)^2-\frac{n}{\sigma^2}\right).$$

On obtient après calcul que $\zeta_n^{RV}=n\log\left(\overline{(X-\mu_0)^2}_n/S_n^2\right)$. *En remarquant qu'on peut décomposer la variance empirique comme*

$$S_n^2=\frac{1}{n}\sum_{i=1}^n(X_i-\mu_0)^2-(\overline{X}_n-\mu_0)^2$$

on obtient directement

$$\zeta_n^{RV}=n\log\left(1+\frac{(\overline{X}_n-\mu_0)^2}{S_n^2}\right).$$

On en déduit la zone de rejet de niveau asymptotique α *de la forme*

$$\left\{\zeta_n^{RV}>q_{1-\alpha}^{\chi_1^2}\right\}\Leftrightarrow\left\{\frac{|\overline{X}_n-\mu_0|}{S_n}>\sqrt{\exp(q_{1-\alpha}^{\chi_1^2}/n)-1}\right\}.$$

On remarque que dans ce cadre gaussien on connaît la loi de $|\overline{X}_n-\mu_0|/S_n$ *: à une transformation croissante près c'est un loi de Student à* $n-1$ *degrés de liberté. On peut ainsi construire un zone de rejet de niveau (non asymptotique)* α *pour le TRV différente de la précédente :*

$$\left\{|\overline{X}_n-\mu_0|>\frac{\sqrt{S_n^2}}{\sqrt{n-1}}q_{1-\alpha/2}^{T_{n-1}}\right\}.$$

On retrouve le test de Student qui est donc un TRV préférable au TRV asymptotique car son risque de premier espèce est exactement α. *En utilisant l'approximation normale sur la loi du* χ^2 *on vérifie bien que les 2 tests coïncident asymptotiquement.*

Chapitre 10

Tests du χ^2

Dans ce chapitre on présente succinctement différents tests fondés sur la statistique du χ^2.

10.1 Tests du χ^2 non paramétriques

On se place dans le cadre du modèle qualitatif à N classes décrit par une variable qualitative X qui prend des valeurs $\{1, \ldots, N\}$ et de loi P telle que

$$P(X = k) = p_k \qquad k \in \{1, \ldots, N\}.$$

Le modèle est dit non paramétrique car la loi P n'appartient pas nécessairement à une loi classique. On sait seulement que le vecteur $\mathbf{p} = (p_1, \ldots, p_N)$ décrit complètement la loi de X et il vérifie $0 \leq p_k \leq 1$ et $\sum_{k=1}^{N} p_k = 1$.

Exemple 10.1.1 *Soit Y un e.a. de $\mathcal{Y}$ de loi P' inconnue quelconque. Le statisticien peut toujours se ramener au cadre précédent en se fixant un entier N et une partition de $\mathcal{Y}$ à N éléments : $\{A_k\}_{1 \leq k \leq N}$. On considère alors la variable discrète X qui vaut k lorsque $Y \in A_k$. Alors par définition $p_k = P'(Y \in A_k)$.*

10.1.1 Test d'adéquation du χ^2 à une loi

Soit $(X_1, \ldots, X_n)$ un échantillon issu du modèle qualitatif à d classes caractérisé par $\mathbf{p} = (p_1, \ldots, p_N)$. On veut savoir si cette échantillon est en adéquation avec le modèle qualitatif à N classes caractérisé par $\mathbf{q} = (q_1, \ldots, q_N)$, $\mathbf{q}$ étant connu (par exemple $\mathbf{q}$ peut correspondre au modèle binomial $\mathcal{B}(N, \theta)$ avec θ connu). On a donc la problématique de test suivante

$$H_0 : \quad \mathbf{p} = \mathbf{q} \qquad \text{contre} \qquad H_1 : \quad \mathbf{p} \neq \mathbf{q}.$$

Définition 10.1.1 *Pour l'adéquation à une loi, on appelle statistique du χ^2 et on note $\hat{\chi}^2_n$ la v.a.*

$$\hat{\chi}^2_n = n \sum_{k=1}^{N} \frac{(\hat{p}_{k,n} - q_k)^2}{q_k}$$

où $\hat{p}_{k,n}$ est la fréquence empirique $\hat{p}_{k,n} = n^{-1} \sum_{i=1}^{n} 1_k(X_i)$.

Remarquons que $\hat{\chi}^2_n$ ne dépend que des observations et de $\mathbf{q}$ connus par le statisticien. Remarquons aussi que sous H_0, la fréquence empirique $\hat{p}_{k,n}$ est la moyenne empirique des $(1_k(X_i)) \sim \mathcal{B}(q_k)$ car $p_k = q_k$. On admettra le résultat suivant

Proposition 10.1.1 *Sous H_0, la statistique du χ_2 vérifie le résultat asymptotique*

$$\hat{\chi}^2_n \xrightarrow{\mathcal{L}} \chi^2_{N-1}.$$

On en déduit une suite de tests ϕ_n de niveau asymptotique α et convergent de zone de rejet

$$R_n = \left\{ \hat{\chi}^2_n > q^{\chi^2_{N-1}}_{1-\alpha} \right\}.$$

Exemple 10.1.2 *Dans sa célèbre expérience, Mendel a étudié l'hérédité de 4 classes distinctes de pois notées $1, \ldots, 4$. Selon que les gènes correspondants soient dominants ou récessifs, il obtient théoriquement une répartition $(9/16, 3/16, 3/16, 1/16)$. Il veut tester avec un niveau d'erreur de 5% la validité de sa théorie génétique sur 556 observations où les effectifs de classes sont $(315, 101, 108, 32)$. On est dans le cadre d'un test d'adéquation à une loi décrite par $\mathbf{q} = (9/16, 3/16, 3/16, 1/16)$. On calcule la statistique $\hat{\chi}^2_n$ correspondante et on obtient*

$$\hat{\chi}^2_n = 556 \left(\frac{(\frac{315}{556} - \frac{9}{16})^2}{\frac{9}{16}} + \frac{(\frac{101}{556} - \frac{3}{16})^2}{\frac{3}{16}} + \frac{(\frac{108}{556} - \frac{3}{16})^2}{\frac{3}{16}} + \frac{(\frac{32}{556} - \frac{1}{16})^2}{\frac{1}{16}} \right) = 0,47$$

que l'on compare à $q^{\chi^2_3}_{0,95} = 0,7815$. Puisque $\hat{\chi}^2_n \leq q^{\chi^2_3}_{0,95}$ il valide sa théorie génétique avec un risque de première espèce asymptotique de 5%. On calcule la p-valeur asymptotique de ce test, c'est à dire le plus petit niveau de risque asymptotique α pour lequel on rejette H_0. Comme on rejette H_0 lorsque $0,47 > q^{\chi^2_3}_{1-\alpha}$ il suffit de trouver le plus petit α vérifiant la relation $\alpha > 1 - F(0,47)$, F étant ici la fonction de répartition d'une χ^2_3. La p-valeur vaut donc $1 - F(0,47) = 0,93$ donc on accepte H_0 mais le test n'est pas significatif. On sait que le test est convergent donc que sa puissance tend vers 1 (son risque de second espèce tend vers 0). Il faudrait calculer la puissance de ce test pour $n = 556$ fixé pour accepter significativement H_0 mais la loi sous H_1 n'est pas spécifiée.

10.1.2 Test d'adéquation du χ^2 à un modèle

On se place toujours dans le cadre non paramétrique où $(X_1, \ldots, X_n)$ est un échantillon issu du modèle qualitatif à N classes caractérisé par $\mathbf{p} = (p_1, \ldots, p_N)$. On veut tester si l'échantillon appartient à un modèle paramétrique (P_θ, Θ) à N classes décrit par $\mathbf{p}(\theta) = (p_1(\theta), \ldots, p_N(\theta))$. On est dans la problématique de test suivante

$$H_0 : \quad \exists \theta \in \Theta \ / \ \mathbf{p} = \mathbf{p}(\theta) \qquad \text{contre} \qquad H_1 : \quad \forall \theta \in \Theta \ / \ \mathbf{p} \neq \mathbf{p}(\theta).$$

Contrairement au cas précédent, sous l'hypothèse nulle le modèle dépend d'un paramètre inconnu θ (par exemple $\mathbf{q}$ peut correspondre au modèle binomial $\mathcal{B}(N, \theta)$ avec $0 < \theta < 1$ inconnu).

On rappelle que pour construire un test d'un niveau (asymptotique) α donné, il suffit de se placer sous l'hypothèse nulle H_0, i.e. dans le modèle paramétrique (P_θ, Θ) ici. Sous H_0, on sait donc estimer le paramètre inconnu θ. On suppose pour simplifier le propos que (P_θ, Θ) est un modèle régulier et identifiable et qu'il existe une unique REV $\hat{\theta}_n$. On sait que $\hat{\theta}_n$ est un "bon" estimateur : elle est asymptotiquement efficace et si elle est sans biais elle est de variance minimale (dans un modèle de la famille exponentielle). On peut ainsi raisonnablement se ramener au test d'adéquation à la loi $\mathbf{q} = \mathbf{p}(\hat{\theta}_n)$:

Définition 10.1.2 *Pour l'adéquation à un modèle, on appelle statistique du χ^2 et on note $\hat{\chi}^2_n$ la v.a.*

$$\hat{\chi}^2_n = n \sum_{k=1}^{N} \frac{(\hat{p}_{k,n} - p_k(\hat{\theta}_n))^2}{p_k(\hat{\theta}_n)}.$$

On rappelle que dans un modèle paramétrique $\Theta \subset \mathbb{R}^d$. On a alors la proposition suivante admise

Proposition 10.1.2 *Si $d < N - 1$ et si la fonction $\theta \to p(\theta)$ est différentiable alors sous H_0 la statistique du χ_2 vérifie le résultat asymptotique*

$$\hat{\chi}^2_n \xrightarrow{\mathcal{L}} \chi^2_{N-d-1}.$$

On en déduit une suite de tests ϕ_n de niveau asymptotique α et convergent de zone de rejet

$$R_n = \left\{ \hat{\chi}^2_n > q_{1-\alpha}^{\chi^2_{N-d-1}} \right\}.$$

Bibliographie

• Livres pour revoir les bases....

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

• Théorie de la mesure et applications aux probabilités

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

• Statistiques inférentielles

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