Préparation à l'agrégation.

Notes de cours

Année 2009 – 2010

Vincent Rivoirard

Le 9 janvier 2010

Chapitre 1

Introduction à la statistique

1.1 Exemples et problématique

Présentons tout d'abord quelques exemples de problèmes statistiques.

Exemple 1.1. *(DDC, chap 2.3) Un projet de loi est soumis à référendum. A cette occasion, on interroge n personnes issues d'une population et on note $X_i = 1$ si la i-ème personne répond "oui"et $X_i = 0$ si elle répond "non". On suppose que les personnes sont choisies au hasard de telle manière que l'on peut supposer que les variables $X_1, \ldots, X_n$ sont i.i.d. de loi de Bernoulli de paramètre $\theta \in]0, 1[$. Au vu des valeurs prises par les variables aléatoires X_i, on cherche à deviner si le projet de loi va être adopté et la proportion des gens de la population qui vont voter "oui". C'est-à-dire que l'on cherche à* **estimer** *θ et à* **tester** *si θ est supérieur à 1/2 ou non.*

Exemple 1.2. *Une entreprise de téléphonie suppose que la distribution du nombre journalier des appels téléphoniques que passent ses clients est une loi de Poisson de paramètre θ inconnu. Cette hypothèse lui permettant de fixer ses tarifs, elle souhaite la* **tester** *à l'aide de la mesure sur une journée du nombre d'appels téléphoniques passés par n clients. Si elle accepte l'hypothèse que la loi est une loi de Poisson elle souhaite également* **estimer** *le paramètre θ de la loi. On suppose donc disposer de l'observation de n variables aléatoires $X_1, \ldots, X_n$ i.i.d. de loi inconnue et on se demande si cette loi est une loi de Poisson.*

Exemple 1.3. *On désire étudier pour la production de blé, l'effet de la variété sur le rendement. On mesure les rendements en blé de $n = 2p$ parcelles de même aire, réparties en deux groupes de taille p suivant la variété représentée. Le rendement observé est l'observation d'une variable aléatoire X_{ij} qui est la somme d'une moyenne théorique m_i, $i \in \{1, 2\}$ et d'une erreur ε_{ij} :*

$$X_{ij} = m_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \overset{i.i.d.}{\sim} \mu, \quad i \in \{1, 2\}, \quad j \in \{1, \ldots, p\},$$

où μ est une loi (connue) centrée de variance σ^2 (connue ou non selon les cas). On cherche donc à ***tester*** *l'hypothèse $m_1 = m_2$ et à* ***estimer*** *la valeur des moyennes théoriques.*

L'objet de la **statistique inférentielle** est de répondre aux problèmes décrits par ces exemples. Il faut noter que comme dans la théorie des probabilités le **hasard** intervient fortement. Mais dans la théorie des probabilités, on suppose la loi connue précisément et on cherche à donner les caractéristiques de la variable qui suit cette loi. L'objectif de la statistique est le contraire : à partir de la connaissance de la variable, que peut-on dire de la loi de cette variable ?

Par ailleurs, pour chacun de ces exemples, nous avons introduit **un modèle statistique** (cette notion sera précisément définie dans le paragraphe suivant). Bien entendu, un modèle est toujours faux. Ce n'est qu'une approximation simple de la réalité que l'on espère pas trop mauvaise. Le choix d'un modèle que l'on doit toujours justifier et critiquer repose sur la connaissance du phénomène étudié, la façon dont une expérience a été menée et sur des représentations graphiques. Un modèle permet l'utilisation des outils mathématiques que l'on va décrire dans la suite de ce cours.

1.2 Modèle statistique

En statistique, on suppose que notre observation (notion définie ci-dessous) est la réalisation d'une variable aléatoire et on se fixe donc un modèle comme décrit précédemment. La loi de la variable n'est pas parfaitement connue, elle dépend d'un paramètre réel, vectoriel ou fonctionnel.

Définition 1.1. *Un **modèle statistique** est la donnée d'un espace mesurable et d'une tribu $(\Omega, \mathcal{A})$ et d'une famille $(\mathbb{P}_\theta)_{\theta\in\Theta}$ de lois de probabilité sur cet espace, i.e. la donnée de $(\Omega, \mathcal{A}, \mathbb{P}_\theta, \theta \in \Theta)$. Quand $\Theta \subset \mathbb{R}^k$, $k \in \mathbb{N}^*$, le modèle est dit **paramétrique**. Sinon, il est dit **non-paramétrique**.*

Définition 1.2. *Une **observation** X est une variable aléatoire à valeurs dans Ω dont la loi appartient à $\{\mathbb{P}_\theta : \quad \theta \in \Theta\}$.*

Remarque 1.1. *L'espace sur lequel X est définie n'a pas d'intérêt statistique et on le confondra avec l'espace d'arrivée (on a le droit !). Si la loi de X est $\mathbb{P}_\theta$, on écrira alors $\mathbb{P}_\theta(X \in A)$ pour tout $A \in \mathcal{A}$.*

La plupart du temps, nos observations auront la structure d'un échantillon.

Définition 1.3. *Pour $n \in \mathbb{N}^*$, un n-**échantillon de loi** $\mathbb{P}$ est la donnée de n variables $X_1, \ldots, X_n$ i.i.d. de loi $\mathbb{P}$.*

Reprenons les exemples précédents.
- Pour l'exemple 1.1 : $\Omega = \{0,1\}^n$, $\mathcal{A}$ est l'ensemble des parties de Ω, $\Theta =]0,1[$, $\mathbb{P}_\theta = \mathcal{B}(\theta)^{\otimes n}$.

- Pour l'exemple 1.2 : $\Omega = \mathbb{N}^n$, $\mathcal{A}$ est l'ensemble des parties de Ω,
$\Theta = \{\theta : \theta =$ fonction de répartition nulle sur $\mathbb{R}_-$ telle que θ est constante sur $[k, k +$

$1[,\ k \in \mathbb{N}\}$, $\mathbb{P}_\theta = \{\text{loi associée à } \theta\}^{\otimes n}$.

- Pour l'exemple 1.3 : Si σ^2 est connue, $\Omega = \mathbb{R}^n$, $\mathcal{A}$ est l'ensemble des boréliens de $\mathbb{R}^n$, $\Theta = \{\theta = (m_1, m_2) \in \mathbb{R}^2\}$, $\mathbb{P}_\theta = \mu_{m_1}^{\otimes p} \otimes \mu_{m_2}^{\otimes p}$ où μ_{m_1} (respectivement μ_{m_2}) est la translatée de la loi μ par m_1 (par respectivement m_2).

Si le modèle est paramétré par θ, on peut espérer dire des choses sur θ à condition que $\theta \longrightarrow \mathbb{P}_\theta$ soit injective. On dit alors que le modèle est **identifiable**.

Dans toute la suite, on suppose donné un modèle statistique identifiable que l'on note $(\Omega, \mathcal{A}, \mathbb{P}_\theta, \theta \in \Theta)$ et on note de la même façon la vraie valeur du paramètre et la valeur courante du paramètre.

1.3 Méthodes d'estimation

Dans le cadre décrit précédemment, l'objectif est d'estimer une quantité dépendant de θ notée $g(\theta)$ où g est une fonction de Θ dans $\mathbb{R}^d$. On s'appuie pour cela sur des estimateurs.

Définition 1.4. *Etant donnée X une observation de loi $\mathbb{P}_\theta$ pour $\theta \in \Theta$, on dit que $T = h(X)$ est un* ***estimateur*** *de la quantité $g(\theta)$ si h est une application mesurable de Ω dans $g(\Theta)$ indépendante de θ.*

C'est la réalisation de cette variable T qui fournit une estimation de $g(\theta)$. Un estimateur peut avoir différentes propriétés qui illustrent son bon comportement pour l'estimation de $g(\theta)$.

Définition 1.5. *On dit que l'estimateur $T = h(X)$ de $g(\theta)$ est*

- ***sans biais*** *si $\forall\ \theta \in \Theta$, $\mathbb{E}_\theta(T) = g(\theta)$ où $\mathbb{E}_\theta$ est l'espérance sous la loi $\mathbb{P}_\theta$ (la quantité $g(\theta) - \mathbb{E}_\theta(T)$ est appelé* **le biais***).*

Quand $X = (X_1, \ldots, X_n)$, on note $T_n = h(X) = h(X_1, \ldots, X_n)$ l'estimateur de $g(\theta)$. Cet estimateur est

- ***asymptotiquement sans biais*** *si $\lim_{n \longrightarrow +\infty} \mathbb{E}_\theta(T_n) = g(\theta)\ \forall\ \theta \in \Theta$,*
- ***consistant*** *si T_n converge en probabilité vers $g(\theta)\ \forall\ \theta \in \Theta$ quand $n \longrightarrow +\infty$,*
- ***fortement consistant*** *si T_n converge presque sûrement vers $g(\theta)\ \forall\ \theta \in \Theta$ quand $n \longrightarrow +\infty$.*

Pour illustrer ce qui précède, reprenons l'Exemple 1.1. On rappelle que pour tout $i \in \{1, \ldots, n\}$, $X_i = 1$ si la i-ème personne vote "oui", 0 si elle vote "non". On a alors $\mathbb{P}_\theta(X_i = 1) = \theta$, $\mathbb{P}_\theta(X_i = 0) = 1 - \theta$. On estime θ par $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. $\overline{X}_n$ est sans biais. De plus, la loi forte des grands nombres montre que $\overline{X}_n$ est fortement consistant, notre estimateur semble donc approprié. Par ailleurs, le TCL montre qu'en loi

$$\frac{\sqrt{n}(\overline{X}_n - \theta)}{\sqrt{\theta(1-\theta)}} \overset{n \to +\infty}{\longrightarrow} \mathcal{N}(0,1).$$

Comme $\theta(1-\theta) \leq 1/4$, on a pour tout $\lambda > 0$, et n proche de l'infini,

$$\mathbb{P}_\theta\left(|\overline{X}_n - \theta| \geq \frac{\lambda}{2\sqrt{n}}\right) \leq \frac{2}{\sqrt{2\pi}} \int_\lambda^{+\infty} \exp(-\frac{x^2}{2}) dx$$

(pour $n\theta \geq 5$, et $n(1-\theta) \geq 5$, l'approximation est convenable). Etant donnée une erreur α fixée, on détermine alors q_α tel que

$$\frac{2}{\sqrt{2\pi}} \int_{q_\alpha}^{+\infty} \exp(-\frac{x^2}{2}) dx = \alpha$$

et on obtient un **intervalle de confiance asymptotique de niveau de confiance** $1-\alpha$ pour θ :

$$I_\alpha = \left[\overline{X}_n - \frac{q_\alpha}{2\sqrt{n}}, \overline{X}_n + \frac{q_\alpha}{2\sqrt{n}}\right].$$

Cet intervalle nous permet de contrôler l'erreur d'estimation. De plus, plus n est grand, plus l'intervalle est de taille réduite et plus l'estimation est précise.
A l'aide de cet intervalle, on peut construire un **test asymptotique de niveau** α de l'hypothèse $H_0 : \theta = \theta_0$ contre $H_1 : \theta \neq \theta_0$. Ainsi, au risque α, on accepte H_0 si $\theta_0 \in I_\alpha$, H_1, sinon.

On sera donc intéressés par des estimateurs sans biais (ou asymptotiquement sans biais) et consistants mais aussi possédant des propriétés de convergence en loi. Décrivons à présent les méthodes les plus classiques pour construire des estimateurs de $g(\theta)$.

1.3.1 Méthode des moments

On suppose donné un n-échantillon $(X_1, \ldots, X_n)$ de loi $\mathbb{P}_\theta$. On veut estimer

$$g(\theta) = \mathbb{E}_\theta[\phi(X_1)] = \int \phi(x) d\mathbb{P}_\theta(x),$$

si $\mathbb{E}_\theta[|\phi(X_1)|] < \infty$. On utilise pour cela la convergence presque sure suivante :

$$\frac{\phi(X_1) + \cdots + \phi(X_n)}{n} \overset{n\to+\infty}{\longrightarrow} g(\theta).$$

Par exemple, pour $A \in \mathcal{A}$, on estime $\mathbb{P}_\theta(X_1 \in A)$ par

$$\frac{1}{n} \sum_{i=1}^{n} 1_{X_i \in A}.$$

On estime $\mathbb{E}_\theta(X_1)$ par

$$\frac{X_1 + \cdots + X_n}{n}$$

et $\mathbb{E}_\theta(X_1^2)$ par

$$\frac{X_1^2 + \cdots + X_n^2}{n}.$$

On peut alors estimer $\mathrm{var}(X_1)$ par

$$\frac{X_1^2 + \cdots + X_n^2}{n} - \left(\frac{X_1 + \cdots + X_n}{n}\right)^2 = \frac{\sum_{i=1}^n (X_1 - \overline{X}_n)^2}{n}.$$

Plus généralement, la méthode des moments consiste à estimer

$$g(\theta) = \psi(g_1(\theta), \ldots, g_p(\theta))$$

où $g_j(\theta) = \mathbb{E}_\theta(\phi_j(X_1))$, $1 \leq j \leq p$, par

$$\psi\left[\frac{\sum_{i=1}^n \phi_1(X_i)}{n}, \ldots, \frac{\sum_{i=1}^n \phi_p(X_i)}{n}\right].$$

Cette méthode permet parfois de déterminer différents estimateurs pour un même paramètre. L'étude des estimateurs obtenus se fait au cas par cas. L'intérêt de cette méthode est qu'il est souvent très facile d'exhiber des estimateurs simples, sans de grosses difficultés de calculs, mais outre que cette méthode est porteuse d'une certaine ambiguïté (plusieurs estimateurs sont possibles parfois même une infinité et alors lequel choisir ?), les estimateurs obtenus peuvent avoir de mauvaises performances.

Exemple 1.4. *L'application de cette méthode à l'Exemple 1.1 fournit l'estimateur* $\overline{X}_n = n^{-1}\sum_{i=1}^n X_i$ *pour estimer* θ*. L'estimateur est sans biais et fortement consistant. Il possède des propriétés de normalité asymptotique.*

Exemple 1.5. *L'application de cette méthode à l'Exemple 1.2 n'est pas possible si on ne connaît pas la distribution du nombre journalier d'appels. Si la distribution est celle d'une loi de Poisson, la méthode fournit l'estimateur* $\overline{X}_n = n^{-1}\sum_{i=1}^n X_i$ *pour estimer* θ*. L'estimateur est sans biais et fortement consistant. Il possède des propriétés de normalité asymptotique.*

Exemple 1.6. *Soit* $i \in \{1, 2\}$*. Pour estimer* m_i *dans l'Exemple 1.3, l'application de cette méthode fournit l'estimateur* $\overline{X}_p^i = p^{-1}\sum_{j=1}^p X_{ij}$*. L'estimateur est sans biais et fortement consistant. Il possède des propriétés de normalité asymptotique.*

Exemple 1.7. *Pour estimer* σ^2 *dans l'Exemple 1.3, la méthode fournit l'estimateur* $S_{ip} = p^{-1}\sum_{j=1}^p X_{ij}^2 - (\overline{X}_p^i)^2 = p^{-1}\sum_{j=1}^p (X_{ij} - \overline{X}_p^i)^2$*. L'estimateur est asymptotiquement sans biais et fortement consistant.*

Pour tous ces exemples, d'autres estimateurs peuvent être obtenus en appliquant la méthode des moments et en calculant les moments d'ordre 2, 3,...

1.3.2 Méthode du maximum de vraisemblance

On suppose donnée une observation X tirée selon une loi $\mathbb{P}_\theta$, $\theta \in \Theta$. On supposera ici que $\mathbb{P}_\theta$ admet une densité par rapport à la mesure de Lebesgue sur $\mathbb{R}^d$, notée f_θ, ou par rapport à la mesure de comptage sur un ensemble dénombrable (et alors $f_\theta(x) = \mathbb{P}_\theta(X = x)$). La méthode du maximum de vraisemblance consiste à estimer $g(\theta) = \theta$ par $\hat{\theta}$ qui maximise la fonction

$$\begin{array}{rcl} \Theta & \longrightarrow & \mathbb{R}^d \\ \theta & \longrightarrow & V_X(\theta) = f_\theta(X) \end{array}$$

Cette fonction est appelée **vraisemblance**. Noter que la vraisemblance est la densité appliquée à l'observation. Quand on dispose d'un n-échantillon $(X_1, \ldots, X_n)$ de loi $\mathbb{P}_\theta$, la vraisemblance est alors

$$V_{X_1,\ldots,X_n}(\theta) = \prod_{i=1}^{n} f_\theta(X_i).$$

L'estimateur du maximum de vraisemblance est alors le point $\theta \in \Theta$ qui maximise la vraisemblance. Cet estimateur est naturel puisqu'il conduit à privilégier la valeur de θ la "plus probable" au vu de l'observation. Il possède en général de bonnes propriétés. Sous certaines conditions, il peut être optimal. L'inconvénient est que ce maximum peut ne pas exister ou ne pas être unique et il peut être difficile à exhiber.

Remarque 1.2. *Plutôt que de maximiser la vraisemblance, on peut de manière équivalente maximiser le logarithme de la vraisemblance (la log-vraisemblance).*

Exemple 1.8. *Pour le modèle de l'Exemple 1.1, la vraisemblance s'écrit :*

$$V_{X_1,\ldots,X_n}(\theta) = \theta^{\sum_{i=1}^n X_i}(1-\theta)^{n-\sum_{i=1}^n X_i}.$$

Le maximum est atteint en un unique point $\hat{\theta}_n = \overline{X}_n$.

Exemple 1.9. *L'application de cette méthode à l'Exemple 1.2 n'est pas possible si on ne connaît pas la distribution du nombre journalier d'appels. Si la distribution est celle d'une loi de Poisson, la vraisemblance s'écrit :*

$$V_{X_1,\ldots,X_n}(\theta) = \exp(-n\theta)\theta^{\sum_{i=1}^n X_i} \prod_{i=1}^{n} (X_i!))^{-1}.$$

Le maximum est atteint en un unique point $\hat{\theta}_n = \overline{X}_n$.

Exemple 1.10. *Pour l'Exemple 1.3, les estimateurs du maximum de vraisemblance ne peuvent pas être déterminés si on ne connaît pas précisément la loi μ. Si on suppose que $\mu = \mathcal{N}(0, \sigma^2)$, les estimateurs du maximum de vraisemblance sont les mêmes que ceux obtenus par la méthode des moments.*

Dans les chapitres suivants, d'autres méthodes moins classiques (estimation par moindres carrés, estimation par quantile,...) seront étudiées.

1.4 Régions de confiance

On estime toujours $g(\theta)$ pour $\theta \in \Theta$ à l'aide d'une observation X.

Définition 1.6. *Soit $\alpha \in [0,1]$. Une* **région de confiance** *de $g(\theta)$ de* **niveau de confiance** $1-\alpha$ *est un ensemble (dépendant de l'observation) $C(X) \subset g(\Theta)$ mesurable par rapport à la tribu sous-jacente au modèle telle que*

$$\forall\, \theta \in \Theta, \quad \mathbb{P}_\theta(g(\theta) \in C(X)) \geq 1-\alpha.$$

Dans le cas d'égalité, on dit que le niveau est **exactement égal** *à $1-\alpha$.*
Si l'observation s'écrit $X = X_n$ (voir précédemment), on parle de ***région de confiance asymptotique de niveau*** $1-\alpha$ *si*

$$\forall\, \theta \in \Theta, \quad \lim_{n\to+\infty} \mathbb{P}_\theta(g(\theta) \in C(X_n)) \geq 1-\alpha.$$

Les valeurs usuelles de α sont 1%, 5% ou 10%. Dans le cas unidimensionnel, la plupart du temps, une région de confiance s'écrit sous la forme d'un intervalle (unilatère ou bilatère). Il faut noter qu'un intervalle de confiance de niveau de confiance 95% a une probabilité au moins égale à 95% de contenir la vraie valeur inconnue $g(\theta)$. A niveau de confiance fixé, une région de confiance est d'autant meilleure qu'elle est de "taille petite". Avant d'aller plus loin, introduisons les quantiles d'une loi de probabilité.

Définition 1.7. *Soit $\alpha \in [0,1]$. On appelle* **quantile d'ordre** α *d'une loi de probabilité $\mathbb{P}$, la quantité*

$$z_\alpha = \inf\left\{x : \quad \mathbb{P}(]-\infty, x]) \geq \alpha\right\}.$$

Par exemple, pour la loi $\mathcal{N}(0,1)$, le quantile d'ordre 97,5% est 1.96, le quantile d'ordre 95% est 1.645. Pour illustrer ce qui précède, considérons l'Exemple 1.3 où pour $i \in \{1,2\}$, on estime m_i à l'aide de $(X_{i1},\ldots,X_{ip})$. On suppose que $\mu = \mathcal{N}(0,1)$. L'estimateur du maximum de vraisemblance de m_i est $\overline{X}_p^i = p^{-1}\sum_{j=1}^p X_{ij}$. Trois intervalles de confiance de niveau de confiance 95% pour l'estimation de m_i sont donné par

- l'intervalle bilatère $[\overline{X}_p^i - 1.96/\sqrt{p}, \overline{X}_p^i + 1.96/\sqrt{p}]$,
- l'intervalle unilatère $[\overline{X}_p^i - 1.645/\sqrt{p}, +\infty[$,
- l'intervalle unilatère $]-\infty, \overline{X}_p^i + 1.645/\sqrt{p}]$.

Pour chacun, on a utilisé le fait que la loi de $\sqrt{p}(\overline{X}_p^i - m_i)$ est indépendante m_i, et ainsi illustré la **méthode du pivot** :

- choix d'un estimateur,
- calcul de sa loi en fonction du paramètre à estimer,
- transformation de cet estimateur pour obtenir une variable aléatoire dont la loi ne dépende plus de ce paramètre.

1.4.1 Intervalles de confiance obtenus par inégalités de probabilité

Application de l'inégalité de Bienaymé-Tchebichev : Rappelons que si X est une variable aléatoire ayant un moment d'ordre 2, alors

$$\forall\, \varepsilon > 0, \quad \mathbb{P}(|X - \mathbb{E}(X)| \geq \varepsilon) \leq \frac{\mathrm{var}(X)}{\varepsilon^2}.$$

Appliquons cette inégalité à l'Exemple 1.1 où on estime θ à l'aide $\overline{X}_n$. On a

$$\forall\, \varepsilon > 0, \quad \mathbb{P}_\theta(|\overline{X}_n - \theta| \geq \varepsilon) \leq \frac{\theta(1-\theta)}{n\varepsilon^2} \leq \frac{1}{4n\varepsilon^2}.$$

On obtient ainsi une région de confiance de niveau $1-\alpha$ en considérant

$$\left[\overline{X}_n - \frac{1}{2\sqrt{n\alpha}}, \overline{X}_n + \frac{1}{2\sqrt{n\alpha}}\right].$$

Pour $\alpha = 5\%$ et $n = 100$, la précision de l'intervalle est 0.22. Il faut noter que la majoration obtenue par l'application de l'inégalité de Bienaymé-Tchebichev n'est pas très précise en particulier si la vraie valeur du paramètre est loin de 1/2.

Application de l'inégalité de Hoeffding (Tsybakov) :

Lemme 1.1. *Soit $(Y_1, \ldots, Y_n)$ une suite de variables indépendantes telles que $\mathbb{E}(Y_i) = 0$ et pour tout i, $a_i \leq Y_i \leq b_i$ p.s., alors*

$$\forall\, \lambda > 0, \quad \mathbb{P}(\sum_{i=1}^n Y_i \geq \lambda) \leq \exp\left(-\frac{2\lambda^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

Preuve : Soit $t > 0$.

$$\begin{aligned}
\mathbb{P}(\sum_{i=1}^n Y_i \geq \lambda) &= \mathbb{E}\left(1_{\sum_{i=1}^n tY_i - t\lambda \geq 0}\right) \\
&\leq \mathbb{E}\left(\exp\left(\sum_{i=1}^n tY_i - t\lambda\right)\right).
\end{aligned}$$

Donc,

$$\begin{aligned}
\mathbb{P}(\sum_{i=1}^n Y_i \geq \lambda) &\leq \exp\left(-\sup_{t>0}\left[t\lambda - \sum_{i=1}^n \log(\mathbb{E}(\exp(tY_i)))\right]\right) \\
&\leq \exp\left(-\sup_{t>0}\left[t\lambda - \frac{t^2}{8}\sum_{i=1}^n (b_i - a_i)^2\right]\right).
\end{aligned}$$

En effet,

$$\begin{aligned}\mathbb{E}(\exp(tY_i)) &\leq \mathbb{E}\left(\frac{b_i - Y_i}{b_i - a_i}\exp(ta_i) + \frac{Y_i - a_i}{b_i - a_i}\exp(tb_i)\right) \\ &\leq \frac{b_i}{b_i - a_i}\exp(ta_i) - \frac{a_i}{b_i - a_i}\exp(tb_i) = \exp(g(u))\end{aligned}$$

où on a posé $u = t(b_i - a_i)$, $s = -a_i/(b_i - a_i)$ et $g(u) = -su + \log(1 - s + s\exp(u))$. On vérifie que $g(0) = g'(0) = 0$ et

$$g''(u) = \frac{(1-s)s\exp(u)}{(1 - s + s\exp(u))^2}.$$

Or $(a+b)^2 = a^2 + b^2 + 2ab \geq 4ab$ donc $g''(u) \leq 1/4$ pour tout u. Par Taylor, pour un $\tau \in [0,1]$,

$$g(u) = \frac{u^2}{2}g''(\tau u) \leq \frac{t^2(b_i - a_i)^2}{8}.$$

On optimise en t avec

$$t = \frac{4\lambda}{\sum_{i=1}^n (b_i - a_i)^2}.$$

□

Appliquons cette inégalité à l'Exemple 1.1 où on estime θ à l'aide de $\overline{X}_n$. On a

$$\forall\, \varepsilon > 0, \quad \mathbb{P}_\theta(|\overline{X}_n - \theta| \geq \varepsilon) \leq 2\exp\left(-2n\varepsilon^2\right).$$

On obtient ainsi une région de confiance de niveau $1 - \alpha$ en considérant

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

Pour $\alpha = 5\%$ et $n = 100$, la précision de l'intervalle est 0.14.

1.4.2 Intervalles de confiance asymptotiques

Supposons que nous cherchions à construire un intervalle de confiance pour un paramètre $g(\theta) \in \mathbb{R}$ à partir d'un échantillon $(X_1, \ldots, X_n)$ de taille n de loi $\mathbb{P}_\theta$ (g est pour l'instant une fonction tout à fait générale). Lorsque nous disposons de suffisamment de données (disons $n \geq 30$) et pour les modèles les plus classiques, le théorème central limite s'avère être le meilleur outil. L'intervalle correspondant est alors un **intervalle de confiance asymptotique**.

Reprenons l'Exemple 1.3 où on suppose $\sigma^2 = 1$. Par application du TCL, pour $i \in \{1, 2\}$,

$$\sqrt{p}\,(\overline{X}_p^i - m_i) \overset{p\to+\infty}{\longrightarrow} \mathcal{N}(0,1) \text{ en loi.}$$

On obtient alors l'intervalle de confiance asymptotique de niveau α suivant pour l'estimation de m_i :

$$\left[\overline{X}_p^i - \frac{\varepsilon_\alpha}{\sqrt{p}}, \overline{X}_p^i + \frac{\varepsilon_\alpha}{\sqrt{p}}\right],$$

où ε_α est le quantile d'ordre $1-\alpha/2$ de la loi $\mathcal{N}(0,1)$.

Pour autant, à ce stade, de multiples problèmes restent en suspens. On considère l'Exemple 1.1. En considérant l'estimateur du maximum de vraisemblance $\overline{X}_n$, le TCL donne :

$$\sqrt{n}(\overline{X}_n - \theta) \overset{n\to+\infty}{\longrightarrow} \mathcal{N}(0, \theta(1-\theta)) \text{ en loi.}$$

La loi limite dépend de θ, ce qui est gênant dans la perspective de construire un intervalle de confiance. Le problème peut se résoudre en remarquant que $\theta(1-\theta) \leq 0.25$. On obtient alors l'intervalle de confiance asymptotique de niveau α suivant :

$$\left[\overline{X}_n - \frac{\varepsilon_\alpha}{2\sqrt{n}}, \overline{X}_n + \frac{\varepsilon_\alpha}{2\sqrt{n}}\right],$$

où ε_α est le quantile d'ordre $1-\alpha/2$ de la loi $N(0,1)$. Pour $\alpha = 5\%$ et $n = 100$, la précision de l'intervalle est 0.10. Dans le cas de l'Exemple 1.2 où on considère $(X_1, \ldots, X_n)$ un échantillon de loi de Poisson de paramètre $\theta > 0$ à estimer, le TCL donne :

$$\sqrt{n}(\overline{X}_n - \theta) \overset{n\to+\infty}{\longrightarrow} \mathcal{N}(0, \theta) \text{ en loi}$$

et des outils plus élaborés doivent être utilisés pour construire un intervalle de confiance dès que l'on ne connaît pas de majorant pour θ. Le **lemme de Slutsky** permet de surmonter certaines difficultés. La **méthode delta** permet d'obtenir des régions de confiances plus intéressantes. Elle permet surtout d'établir la normalité asymptotique de nombreux estimateurs pour l'estimation de $g(\theta)$. Ces résultats sont présentés et démontrés dans les sections suivantes.

Lemme 1.2. *(***Slutsky***) Soient $(X_n)_{n\in\mathbb{N}}$ et $(Y_n)_{n\in\mathbb{N}}$ deux suites de vecteurs de $\mathbb{R}^d$, $d \in \mathbb{N}^*$ et $\mathbb{R}^p$, $p \in \mathbb{N}^*$, respectivement tels que*

- *$X_n \overset{n\to+\infty}{\longrightarrow} X$ en loi où X est un vecteur aléatoire,*
- *$Y_n \overset{n\to+\infty}{\longrightarrow} c$ en probabilité où c est un vecteur déterministe,*

alors $(X_n, Y_n) \overset{n\to+\infty}{\longrightarrow} (X, c)$ en loi.

Preuve : Soit $f : \mathbb{R}^d \times \mathbb{R}^p \longrightarrow \mathbb{R}$ une fonction uniformément continue bornée.

$$\begin{aligned}
&|\mathbb{E}f(X_n, Y_n) - \mathbb{E}f(X, c)| \\
&\qquad \leq |\mathbb{E}f(X_n, c) - \mathbb{E}f(X, c)| + |\mathbb{E}f(X_n, Y_n) - \mathbb{E}f(X_n, c)| \\
&\qquad \leq |\mathbb{E}f(X_n, c) - \mathbb{E}f(X, c)| + |\mathbb{E}(f(X_n, Y_n) - f(X_n, c))1_{\|Y_n - c\| \leq \delta}| + 2\|f\|_\infty \mathbb{P}(\|Y_n - c\| > \delta),
\end{aligned}$$

qui tend bien vers 0 compte tenu des hypothèses. On conclut par le lemme du Portmanteau et la densité pour la norme sup des fonctions uniformément continues bornées dans

l'ensemble $\{1_U : U \text{ ouvert}\}$. □

En reprenant l'Exemple 1.2, on obtient :

$$\frac{\sqrt{n}(\overline{X}_n - \theta)}{\sqrt{\overline{X}_n}} \overset{n\to+\infty}{\longrightarrow} \mathcal{N}(0,1) \text{ en loi}$$

mais aussi avec

$$s_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$$

qui est un estimateur consistant de la variance

$$\frac{\sqrt{n}(\overline{X}_n - \theta)}{s_n} \overset{n\to+\infty}{\longrightarrow} \mathcal{N}(0,1) \text{ en loi.}$$

Et ainsi de suite... Lequel choisir ? La méthode delta va fournir une alternative en offrant un estimateur naturel et dans le cas de l'Exemple 1.2 bien meilleur d'un point de vue numérique. Intuitivement, ce dernier résultat s'explique bien. En effet, le lemme de Slutsky qui permet d'obtenir la méthode **plug in** décrite précédemment estime les paramètres gênants alors que notre but est d'essayer de contrôler la variance de $\overline{X}_n$. On devine que cet aléa que l'on ajoute ne peut que modifier dans la mauvaise direction la variance de notre estimateur. Cet exemple pourra être relié à l'exemple étudié dans le cours sur le modèle linéaire gaussien quand on estime la moyenne selon que la variance est connue ou non (comparaison des queues des lois $\mathcal{N}(0,1)$ et de Student). Par ailleurs, en utilisant le lemme de Slutsky, la longueur de l'intervalle de confiance obtenu dépend des observations. La encore, la méthode delta pourra remédier à ce problème.

Pour énoncer le théorème associé à la **méthode delta**, fixons les notations. On suppose que $g : \mathbb{R}^d \longrightarrow \mathbb{R}^p$ pour d et p dans $\mathbb{N}^*$. On note indistinctement $\|.\|$ la norme euclidienne dans $\mathbb{R}^d$ ou $\mathbb{R}^p$.

Théorème 1.1. *On se donne une suite $(U_n)_{n\in\mathbb{N}}$ de vecteurs aléatoires de $\mathbb{R}^d$ et une suite déterministe $(a_n)_{n\in\mathbb{N}}$ telle que $a_n \overset{n\to+\infty}{\longrightarrow} +\infty$. On suppose :*

– $a_n(U_n - U) \overset{n\to+\infty}{\longrightarrow} V$ en loi pour V et U deux vecteurs de $\mathbb{R}^d$ (U est non-aléatoire).

– g est une fonction différentiable en U dont la différentielle est notée $Dg(U) \in \mathbb{R}^{p\times d}$.

Alors

$$a_n(g(U_n) - g(U)) \overset{n\to+\infty}{\longrightarrow} Dg(U) \times V \text{ en loi.}$$

Preuve : Par le lemme de Slutsky, il suffit de montrer que

$$a_n\left[g(U_n) - g(U) - Dg(U)(U_n - U)\right] \overset{n\to+\infty}{\longrightarrow} 0 \text{ en proba.}$$

Soit $\delta > 0$. Soient $M > 0$ et $\varepsilon = M^{-1}\delta$. Par définition, il existe $\delta' > 0$ tel que pour tout $U' \in \mathbb{R}^d$,

$$\|U' - U\| \leq \delta' \Rightarrow \|g(U) - g(U') - Dg(U)(U' - U)\| \leq \varepsilon\|U' - U\|.$$

On a pour n assez grand :

$$\begin{aligned}
&\mathbb{P}(\|a_n(g(U_n)-g(U)-Dg(U)(U_n-U))\|>\delta)\\
&\qquad\leq \mathbb{P}(\|a_n(g(U_n)-g(U)-Dg(U)(U_n-U))\|>\delta, \|U_n-U\|\leq\delta')+\mathbb{P}(\|U_n-U\|>\delta')\\
&\qquad\leq \mathbb{P}(a_n\varepsilon\|U_n-U\|>\delta)+\mathbb{P}(\|U_n-U\|>\delta')\\
&\qquad\leq 2\mathbb{P}(a_n\|U_n-U\|\geq M)\\
&\qquad= 2(1-\mathbb{P}(a_n\|U_n-U\|<M))
\end{aligned}$$

Rappelons que par le lemme du Portmanteau, $a_n(U_n-U)\overset{n\to+\infty}{\longrightarrow} V$ en loi $\Leftrightarrow \liminf_{n\to+\infty}\mathbb{P}(a_n(U_n-U)\in O)\geq\mathbb{P}(V\in O)$ pour tout ouvert O. Donc, on conclut que

$$\limsup_{n\to+\infty}\mathbb{P}(\|a_n(g(U_n)-g(U)-Dg(U)(U_n-U))\|>\delta)\leq 2(1-\mathbb{P}(\|V\|<M)),$$

ceci pour tout M, ce qui achève la preuve. □

Remarque 1.3. *Noter que l'on a démontré au passage que si* $a_n\overset{n\to+\infty}{\longrightarrow}+\infty$, *alors*

$$a_n(U_n-U)\overset{n\to+\infty}{\longrightarrow} V \text{ en loi} \Rightarrow U_n\overset{n\to+\infty}{\longrightarrow} U \text{ en proba.}$$

On a alors :

Corollaire 1.1. *On se donne* $(X_1,\dots,X_n)$ *un échantillon de vecteurs de* $\mathbb{R}^d$. *On suppose que* X_1 *est de carré intégrable et on note* μ *la moyenne de* X_1 *et* Σ *sa matrice de variance-covariance. Si* g *est une fonction différentiable en* U *dont la différentielle est notée* $Dg(U)\in\mathbb{R}^{p\times d}$, *alors*

$$\sqrt{n}(g(\overline{X}_n)-g(\mu))\overset{n\to+\infty}{\longrightarrow}\mathcal{N}(0,Dg(U)\Sigma Dg(U)^*) \text{ en loi.}$$

Preuve : On applique le TCL vectoriel et le théorème précédent. □

Revenons au problème initial en considérant à nouveau l'Exemple 1.2. Quelle fonction g va-t-on choisir ? Naturellement, on considère la fonction g qui permet d'obtenir une loi limite indépendante de θ, donc $g(\theta)=c_1\sqrt{\theta}+c_2$, c_1 et c_2 deux constantes indépendantes de θ. On a alors avec $g(\theta)=\sqrt{\theta}$:

$$\sqrt{n}(\overline{X}_n^{1/2}-\theta^{1/2})\overset{n\to+\infty}{\longrightarrow}\mathcal{N}(0,0.25).$$

On obtient l'intervalle de confiance pour $\theta^{1/2}$:

$$I_{n,\alpha}=\left[\overline{X}_n^{1/2}-\frac{\varepsilon_\alpha}{2\sqrt{n}},\overline{X}_n^{1/2}+\frac{\varepsilon_\alpha}{2\sqrt{n}}\right],$$

où ε_α est le quantile d'ordre $1-\alpha/2$ de la loi $\mathcal{N}(0,1)$ et l'intervalle de confiance pour θ :

$$g^{-1}(I_{n,\alpha}) = \left[\left(\overline{X}_n^{1/2} - \frac{\varepsilon_\alpha}{2\sqrt{n}} \right)^2, \left(\overline{X}_n^{1/2} + \frac{\varepsilon_\alpha}{2\sqrt{n}} \right)^2 \right]$$

A noter qu'on obtient un intervalle de confiance pour $\theta^{1/2}$ de taille indépendante des données (**stabilisation de la variance**). Cette remarque peut fréquemment être mise à profit dans les problèmes de tests. La méthode delta permet surtout d'obtenir la normalité asymptotique d'estimateurs pour l'estimation de $g(\theta)$ (vitesse $\sqrt{n}$) pour une large classe de fonctions g. Ce résultat sera essentiel quand on étudiera l'optimalité de méthodes d'estimation.

Reprenons l'Exemple 1.3 où on cherche à estimer σ^2 à l'aide de l'estimateur obtenu par la méthode des moments :

$$S_{1p} = p^{-1} \sum_{j=1}^{p} X_{1j}^2 - (\overline{X}_p^1)^2 = p^{-1} \sum_{j=1}^{p} (X_{1j} - \overline{X}_p^1)^2.$$

Posons

$$\forall\, j \in \{1, \ldots, p\}, \quad Z_{1j} = \begin{pmatrix} X_{1j}^2 \\ X_{1j} \end{pmatrix}.$$

Le TCL vectoriel donne

$$\sqrt{p} \left(\frac{1}{p} \sum_{j=1}^{p} Z_{1j} - \begin{pmatrix} \mathbb{E}(X_{11}^2) \\ \mathbb{E}(X_{11}) \end{pmatrix} \right) \overset{p \to +\infty}{\longrightarrow} \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Gamma \right)$$

en loi avec

$$\Gamma = \begin{pmatrix} \mathbb{E}(X_{11}^4) - (\mathbb{E}(X_{11}^2))^2 & \mathbb{E}(X_{11}^3) - \mathbb{E}(X_{11})\mathbb{E}(X_{11}^2) \\ \mathbb{E}(X_{11}^3) - \mathbb{E}(X_{11})\mathbb{E}(X_{11}^2) & \mathbb{E}(X_{11}^2) - (\mathbb{E}(X_{11}))^2 \end{pmatrix}.$$

Si on considère

$$\begin{array}{rccc} g: & \mathbb{R}^2 & \longrightarrow & \mathbb{R} \\ & \begin{pmatrix} x \\ y \end{pmatrix} & \to & x - y^2, \end{array}$$

on obtient

$$\sqrt{p}(S_{1p} - \sigma^2) \overset{p \to +\infty}{\longrightarrow} \mathcal{N}(0, \Sigma) \tag{1.1}$$

où

$$\Sigma = (1, -2\mathbb{E}(X_{11})) \begin{pmatrix} \mathbb{E}(X_{11}^4) - (\mathbb{E}(X_{11}^2))^2 & \mathbb{E}(X_{11}^3) - \mathbb{E}(X_{11})\mathbb{E}(X_{11}^2) \\ \mathbb{E}(X_{11}^3) - \mathbb{E}(X_{11})\mathbb{E}(X_{11}^2) & \mathbb{E}(X_{11}^2) - (\mathbb{E}(X_{11}))^2 \end{pmatrix} \begin{pmatrix} 1 \\ -2\mathbb{E}(X_{11}) \end{pmatrix}.$$

Une nouvelle application de la méthode des moments et du lemme de Slutsky permet de construire un intervalle de confiance asymptotique pour σ^2.

1.5 Tests

1.5.1 Généralités

Dans le cadre de notre modèle statistique $(\Omega, \mathcal{A}, \mathbb{P}_\theta, \theta \in \Theta)$, on se donne Θ_0 et Θ_1 disjoints tels que $\Theta_0 \cup \Theta_1 = \Theta$. Au vu d'une observation X on veut décider si $\theta \in \Theta_0$ ou pas (et alors $\theta \in \Theta_1$). On note

- **l'hypothèse nulle** $H_0 : \theta \in \Theta_0$,
- **l'hypothèse alternative** $H_1 : \theta \in \Theta_1$.

Pour $i \in \{0, 1\}$, H_i et dite **simple** si Θ_i est réduit à un singleton, **composite** sinon. Nous verrons que H_0 et H_1 ne jouent pas un rôle symétrique. Pour résoudre notre problème de décision, on s'appuie sur la définition suivante.

Définition 1.8. *On appelle* ***test de l'hypothèse*** H_0 ***contre l'hypothèse*** H_1 *toute fonction mesurable de l'observation X, notée $\phi(X)$ à valeurs dans $\{0, 1\}$. Si $\phi(X) = 0$, on accepte H_0. Si $\phi(X) = 1$, on rejette H_0 et on accepte H_1.*

Remarque 1.4. *Ainsi $\phi(X)$ s'écrit $\phi(X) = 1_{X \in R}$. R est appelée la* ***région de rejet de*** *H_0, R^c est appelée la* ***région d'acceptation de*** *H_0.*

Donnons à présent quelques définitions.

Définition 1.9. *On définit*

- ***le risque de première espèce :***

$$\begin{aligned} \Theta_0 &\longrightarrow [0, 1] \\ \theta &\longrightarrow \mathbb{P}_\theta(\phi(X) = 1) \end{aligned}$$

- ***le risque de seconde espèce :***

$$\begin{aligned} \Theta_1 &\longrightarrow [0, 1] \\ \theta &\longrightarrow \mathbb{P}_\theta(\phi(X) = 0) \end{aligned}$$

- ***la puissance du test :***

$$\begin{aligned} \Theta_1 &\longrightarrow [0, 1] \\ \theta &\longrightarrow \mathbb{P}_\theta(\phi(X) = 1) \end{aligned}$$

- ***la taille du test :*** $\sup_{\theta \in \Theta_0} \mathbb{P}_\theta(\phi(X) = 1)$. *On dit qu'un test est* ***de niveau*** α *si sa taille est inférieure ou égale à α.*

Bien entendu, on cherche des tests de faibles risques. Pour illustrer ces notions, reprenons l'Exemple 1.3 avec la première variété. L'exploitant agricole a reçu l'assurance du vendeur de cette variété que le rendement serait au moins égal à 1. Il décide de tester cette affirmation. En faisant l'hypothèse que $\mu = \mathcal{N}(0, 1)$, il teste donc $H_0 : m_1 \geq 1$ contre $H_1 : m_1 < 1$. On a donc $\Theta_0 = [1, +\infty[$ et $\Theta_1 =]-\infty, 1[$. Il décide de considérer le test :

$$\phi(X) = 1_{\overline{X}_p^1 < 1}$$

avec $X = (X_{11}, \ldots, X_{1p})$, et $\overline{X}_p^1 = p^{-1} \sum_{j=1}^p X_{1j}$. En notant F la fonction de répartition de la loi $\mathcal{N}(0,1)$, Le risque de première espèce est la fonction :

$$\begin{array}{rcl} \Theta_0 & \longrightarrow & [0,1] \\ \theta & \longrightarrow & F(\sqrt{p}(1-\theta)) \end{array}$$

le risque de seconde espèce est la fonction :

$$\begin{array}{rcl} \Theta_1 & \longrightarrow & [0,1] \\ \theta & \longrightarrow & 1 - F(\sqrt{p}(1-\theta)) \end{array}$$

la puissance du test est la fonction :

$$\begin{array}{rcl} \Theta_1 & \longrightarrow & [0,1] \\ \theta & \longrightarrow & F(\sqrt{p}(1-\theta)) \end{array}$$

La taille du test est 0.5. Dans cet exemple, $\overline{X}_p^1$ est appelée **la statistique de test** et par abus, $\{\overline{X}_p^1 < 1\}$ la région de rejet. Présentée telle qu'elle, la théorie des tests semblent faire jouer un rôle symétrique à H_0 et H_1, il n'en est rien. Explicitons cette dissymétrie.

- Quand on construit un test, on veut mesurer l'adéquation de l'hypothèse nulle avec les observations. On vérifie si le modèle associé à H_0 n'est pas en contradiction avec les données. S'il y a un désaccord grave, on rejette H_0, sinon et peut-être faute de mieux, on conserve H_0 (c'est moins que de dire qu'il y a accord avec les données).
- Les deux risques ne jouent pas le même rôle. Si on se fixe un niveau α, α peut-etre vu comme le risque maximal que l'on accepte de prendre en rejetant H_0 à tort. Le risque de première espèce est privilégié. On fixe d'abord α puis on cherche des tests de niveau α les plus puissants possibles.

Ces considérations peuvent nous aider à déterminer H_0 et H_1. On prendra pour H_0 :

- une hypothèse communément établie,
- une hypothèse de prudence (critère de coût, de sécurité,...),
- la seule facile à formuler.

La démarche est la suivante :

1. Choix de H_0 et H_1
2. Détermination de la statistique de test $T(X)$ (on doit en connaître la loi sous H_0)
3. Allure de la zone de rejet en fonction de la forme de H_1
4. Calcul de la zone de rejet en fonction du niveau α préalablement fixé
5. Calcul de la valeur expérimentale T_{obs} de $T(X)$ sur les données
6. Conclusion

On peut compléter ces étapes en calculant la puissance. Raffinons l'exemple précédent en donnant des valeurs numériques. On suppose que $p = 4$ et que l'on observe $X_{11} = 0.15$, $X_{12} = 0.45$, $X_{13} = 0.9$ et $X_{14} = 0.5$. L'exploitant se demande si ces valeurs ne contredisent

pas les affirmations du vendeur. Envisageant de lui intenter un procès, il décide de faire un test basé sur la statistique de test précédente $\overline{X}_p^1$ avec une région de rejet de la même forme, mais pour des raisons évidentes de coût, il se fixe un niveau pas trop grand : $\alpha = 10\%$. Son test est

$$\phi(X) = 1_{\overline{X}_p^1 < k_{\alpha,p}}.$$

et il faut donc $\sqrt{p}(k_{\alpha,p} - 1) < -1.28$. Et avec $k_{\alpha,p} = 1 - \frac{1.28}{\sqrt{p}} \approx 0.36$, la taille du test vaut 10%. On calcule $\overline{X}_p^1 = 0.5$ donc on accepte H_0. La puissance du test est la fonction :

$$\begin{array}{rcl} \Theta_1 & \longrightarrow & [0,1] \\ \theta & \longrightarrow & F(\sqrt{p}(0.36 - \theta)). \end{array}$$

Une variante basée sur la **p-valeur** est la suivante :

1. Choix de H_0 et H_1
2. Détermination de la statistique de test $T(X)$ (on doit en connaître la loi sous H_0)
3. Allure de la zone de rejet en fonction de la forme de H_1
4. Calcul de la valeur expérimentale T_{obs} de $T(X)$ sur les données
5. Calcul de la p**-valeur** α_0 en injectant la valeur de T_{obs} dans la constante intervenant dans la zone de rejet. α_0 est la plus petite valeur du niveau du test qui autorise le rejet de H_0 (α_0 dépend des observations).
6. Si l'examinateur vous donne la valeur de α, conclusion en comparant α à α_0

Reprenons l'exemple précédent pour obtenir la calcul de la p-valeur :

$$\alpha_0 = \mathbb{P}_{\theta=1}(\overline{X}_p^1 < 0.5) = \mathbb{P}_{\theta=1}(\sqrt{p}(\overline{X}_p^1 - 1) < \sqrt{p}(0.5-1)) = \mathbb{P}(Z < -1) = 15.9\%, \quad Z \sim \mathcal{N}(0,1).$$

Il faut noter la dualité entre le problème des tests et celui de la construction des régions de confiance.

- Si on dispose d'une région de confiance $C(X)$ de niveau $1-\alpha$ pour l'estimation de θ alors pour tout θ_0 le test
$$\phi(X) = 1_{\{\theta_0 \notin C(X)\}}$$
est un test de niveau α pour le test de $H_0 : \theta = \theta_0$ contre $H_1 : \theta \neq \theta_0$.
- Réciproquement, si pour tout θ_0, on dispose d'un test $\phi_{\theta_0}(X)$ de $\theta = \theta_0$ contre $\theta \neq \theta_0$ alors
$$C(X) = \{\theta : \quad \phi_\theta(X) = 0\}$$
est une région de confiance de niveau $1-\alpha$ pour l'estimation de θ.

Illustrons le premier point à l'aide de (1.1). On se donne σ_0 et on veut tester au niveau α $H_0 : \sigma = \sigma_0$ contre $H_1 : \sigma \neq \sigma_0$. On suppose qu'il existe un estimateur $\tilde{\Sigma}$ de Σ consistant. En notant ε_α le quantile d'ordre $1 - \alpha/2$ de la loi $\mathcal{N}(0,1)$, une région de confiance asymptotique de niveau $1-\alpha$ pour σ^2 est donnée par

$$\left[S_{1p} - \frac{\varepsilon_\alpha \sqrt{\tilde{\Sigma}}}{\sqrt{p}}, S_{1p} + \frac{\varepsilon_\alpha \sqrt{\tilde{\Sigma}}}{\sqrt{p}} \right].$$

Asymptotiquement, on accepte donc H_0 si et seulement si

$$\sigma_0^2 \in \left[S_{1p} - \frac{\varepsilon_\alpha \sqrt{\tilde{\Sigma}}}{\sqrt{p}}, S_{1p} + \frac{\varepsilon_\alpha \sqrt{\tilde{\Sigma}}}{\sqrt{p}} \right] \iff \frac{\sqrt{p}|S_{1p} - \sigma_0^2|}{\sqrt{\tilde{\Sigma}}} \leq \varepsilon_\alpha.$$

On distingue en général 4 types de tests.

1. Les **tests de conformité** où on teste si l'observation X est issue d'une population caractérisée par une valeur précise d'un paramètre.
   Exemple : si on note m la moyenne de X, $H_0 : m = 0$ contre $H_1 : m \neq 0$.
2. Les **tests d'ajustement à une loi ou à une famille de loi** où on teste si l'observation X possède une loi ou un type de loi.
   Exemple : En notant μ la loi de X, on teste : $H_0 : \mu = \mathcal{N}(0,1)$ contre $H_1 : \mu \neq \mathcal{N}(0,1)$.
   Exemple : H_0 : il existe m et σ^2 tels que $\mu = \mathcal{N}(m, \sigma^2)$ contre H_1 : il n'existe pas m et σ^2 tels que $\mu = \mathcal{N}(m, \sigma^2)$.
3. Les **tests d'homogénéité** où on teste si 2 observations X et Y sont issues d'une même population.
   Exemple : si on note m la moyenne de X et m' la moyenne de Y, $H_0 : m = m'$ contre $H_1 : m \neq m'$.
   Exemple : En notant μ la loi de X et ν la loi de Y, on teste $H_0 : \mu = \nu$ contre $H_1 : \mu \neq \nu$.
4. Les **tests d'indépendance** où on teste si 2 échantillons X et Y sont de lois indépendantes ou non.

Remarque 1.5. *Le test d'indépendance peut bien-sûr être ré-écrit comme un test d'ajustement à l'ensemble des lois-produit. Le test d'homogénéité peut être vu comme un test d'indépendance. En effet, si $X = (X_1, \ldots, X_n)$ et $Y = (Y_1, \ldots, Y_n)$, tester l'homogénéité de X et Y revient à tester l'indépendance de $((X_1, 0), \ldots, (X_n, 0), (Y_1, 1), \ldots, (Y_n, 1))$.*

L'exemple étudié tout au long de cette section est un exemple de test de conformité. Donnons à présent un exemple de test d'homogénéité de niveau α en testant l'hypothèse $H_0 : m_1 = m_2$ contre $H_1 : m_1 \neq m_2$ dans le cadre de l'Exemple 1.3. On suppose toujours que $\mu = \mathcal{N}(0,1)$ et que $\sigma^2 = 1$. Très naturellement, on s'appuie sur la statistique de test

$$\overline{X}_p = \overline{X}_p^1 - \overline{X}_p^2 = \frac{1}{p} \sum_{j=1}^{p} (X_{1j} - X_{2j}) \sim \mathcal{N}\left(0, \frac{2}{p}\right) \text{ sous } H_0.$$

On rejette H_0 dès que

$$\frac{|\sum_{j=1}^{p} (X_{1j} - X_{2j})|}{\sqrt{2p}} > \varepsilon_\alpha,$$

où ε_α est le quantile d'ordre $1 - \alpha/2$ de la loi $\mathcal{N}(0,1)$. Bien entendu, on peut aussi considérer l'alternative $H_1 : m_1 > m_2$ ou l'alternative $H_1 : m_1 < m_2$. On prend alors la

même statistique de test mais sans les valeurs absolues.

Un test peut posséder diverses propriétés que nous ne détaillerons pas ici (la robustesse, être sans biais, libre et dans un cadre asymptotique, la consistance). Néanmoins, pour comparer deux tests, on utilisera la définition suivante.

Définition 1.10. *Si $\phi(X)$ et $\phi'(X)$ sont deux tests de niveau α, on dit que $\phi(X)$ est* ***uniformément plus puissant*** *que $\phi'(X)$ si*

$$\forall\, \theta \in \Theta_1, \quad \mathbb{P}_\theta(\phi(X) = 1) \geq \mathbb{P}_\theta(\phi'(X) = 1).$$

$\phi(X)$ est dit ***UPP***(α) *s'il est uniformément plus puissant que tout test de niveau α.*

Un test UPP(α) n'existe pas toujours mais nous allons voir dans la section suivante dans quel cadre on peut en construire un.

1.5.2 Tests de rapport de vraisemblance

Dans cette section, nous allons nous arrêter sur la construction d'un test intuitif : **le test de rapport de vraisemblance**. Par convention, on supposera $0/0 = 0$. On suppose que le modèle statistique est dominé par une mesure μ qui est la mesure de Lebesgue sur $\mathbb{R}^d$ ou la mesure de comptage sur un ensemble dénombrable. On notera f_θ, la densité par rapport à cette mesure μ. La statistique du test de rapport de vraisemblance s'écrit :

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

où $V_X(\theta) = f_\theta(X)$ est la vraisemblance du modèle au point θ et le test de rapport de vraisemblance s'écrit sous la forme :

$$\phi(X) = 1_{T(X)>k_\alpha}.$$

On peut l'écrire en général de manière plus simple. Cette construction est intuitive puisque l'interprétation de la vraisemblance conduit naturellement à rejeter H_0 si le numérateur est très supérieur au dénominateur.

Exemple : Reprenons l'Exemple 1.1. On a vu que la vraisemblance s'écrit :

$$\forall\, \theta \in \Theta, \quad V_X(\theta) = \theta^{\sum_{i=1}^n X_i}(1-\theta)^{n-\sum_{i=1}^n X_i}.$$

On veut tester $H_0 : \theta \leq 1/2$ contre $H_1 : \theta > 1/2$. La statistique de test s'écrit :

$$T(X) \;=\; \begin{cases} \left(\frac{\sum_{i=1}^n X_i/n}{1/2}\right)^{\sum_{i=1}^n X_i} \left(\frac{1-\sum_{i=1}^n X_i/n}{1-1/2}\right)^{n-\sum_{i=1}^n X_i} & \text{si } \sum_{i=1}^n X_i/n > 1/2, \\ \left(\frac{1/2}{\sum_{i=1}^n X_i/n}\right)^{\sum_{i=1}^n X_i} \left(\frac{1-1/2}{1-\sum_{i=1}^n X_i/n}\right)^{n-\sum_{i=1}^n X_i} & \text{si } \sum_{i=1}^n X_i/n \leq 1/2. \end{cases}$$

La fonction $u \to u\log(u) + (1-u)\log(1-u)$ étant décroissante sur $]0, 1/2]$ et croissante sur $[1/2, 1[$ le test de rapport de vraisemblance s'écrit sous la forme :

$$\phi(X) = 1_{\sum_{i=1}^n X_i > x_{n,\alpha}}.$$

Le **lemme de Neyman-Pearson** montre l'optimalité des tests de rapport de vraisemblance dans le cadre de tests d'hypothèses simples.

Lemme 1.3. *Dans le cadre du problème du test de $H_0 : \theta = \theta_0$ contre $H_1 : \theta = \theta_1$, si $\phi(X)$ est le test de rapport de vraisemblance de taille $\alpha > 0$, c'est-à-dire :*

$$\phi(X) = 1_{L(X,\theta_0,\theta_1) > k_\alpha}, \quad L(X, \theta_0, \theta_1) = \frac{V_X(\theta_1)}{V_X(\theta_0)} = \frac{f_{\theta_1}(X)}{f_{\theta_0}(X)},$$

et

$$\mathbb{E}_{\theta_0}(\phi(X)) = \mathbb{P}_{\theta_0}(\phi(X) = 1) = \alpha > 0,$$

alors $\phi(X)$ est un test UPP(α).

Preuve : Soit $\psi(X)$ un test de niveau α. Donc $\mathbb{E}_{\theta_0}(\psi(X)) = \mathbb{P}_{\theta_0}(\psi(X) = 1) \leq \alpha$. On veut montrer que

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

Observons que $\alpha > 0 \Rightarrow k_\alpha < +\infty \Rightarrow \phi(X) = 1$ si $f_{\theta_1}(X) > 0$ et $f_{\theta_0}(X) = 0$.

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

□

Remarque 1.6. *L'hypothèse $\alpha > 0$ peut être omise à condition de considérer $\phi(X) = 1_{V_X(\theta_1) > k_\alpha V_X(\theta_0)}$. Mais dans ce cas là, k_α peut être égal à $+\infty$.*

Remarque 1.7. *Dans notre cadre et pour tout $\alpha \in [0, 1]$, il n'existe pas forcément un test de taille α. On contourne cette difficulté en considérant les tests randomisés.*

Pour illuster le résultat donné par le lemme de Neyman-Pearson, on peut considérer l'Exemple 1.3, où on teste $H_0 : m_1 = 1$ contre $H_1 : m_1 = 0$ sous l'hypothèse $\mu = \mathcal{N}(0, 1)$ et $\sigma^2 = 1$. En notant $X = (X_{11}, \ldots, X_{1p})$, et $\overline{X}_p^1 = p^{-1}\sum_{j=1}^p X_{1j}$, le test UPP($\alpha$) s'écrit

$$\phi(X) = 1_{\overline{X}_p^1 < 1 - \frac{\varepsilon_\alpha}{\sqrt{p}}}$$

où ε_α est le quantile d'ordre $1-\alpha$ de la loi $\mathcal{N}(0, 1)$.

Chapitre 2

Vecteurs gaussiens - Tests du $\chi 2$

2.1 Introduction

2.1.1 Définitions

Rappelons la définition des variables aléatoires gaussiennes réelles.

Définition 2.1. *On définit :*

- *Une variable aléatoire réelle Z est dite* **gaussienne centrée réduite** *si elle admet pour densité par rapport à la mesure de Lebesgue sur $\mathbb{R}$ la fonction :*

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

  On note $Z \sim \mathcal{N}(0,1)$.
- *Une variable aléatoire réelle X est dit* **gaussienne** *s'il existe $(\mu, \sigma) \in \mathbb{R} \times \mathbb{R}_+$ et $Z \sim \mathcal{N}(0,1)$ tels que $X = \mu + \sigma Z$. La densité de X est*

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

  On note $X \sim \mathcal{N}(\mu, \sigma^2)$. Quand $\sigma = 0$, on dit que X est une variable gaussienne dégénérée.

Une variable gaussienne est caractérisée par sa fonction caractéristique donnée par la proposition suivante.

Théorème 2.1. *La fonction caractéristique de $X \sim \mathcal{N}(\mu, \sigma^2)$ est donnée par*

$$\forall\, t \in \mathbb{R}, \quad \phi_X(t) = \exp\left(it\mu - \frac{\sigma^2 t^2}{2}\right).$$

Preuve : ϕ_X se calcule à l'aide de ϕ_Z où $Z \sim \mathcal{N}(0,1)$ et on montre que

$$\forall\, t \in \mathbb{R}, \quad \phi_Z(t) = -t\phi_Z'(t).$$

□

Introduisons à présent les vecteurs gaussiens.

Définition 2.2. *Un vecteur aléatoire X à valeurs dans $\mathbb{R}^d$ est dit* **gaussien** *si toute combinaison linéaire de ses composantes est une variable aléatoire gaussienne.*
Si $X = (X_1, \ldots, X_d)^*$ est un vecteur gaussien, on définit son **vecteur moyenne** $\mathbb{E}(X)$ *par*

$$\mathbb{E}(X) = (\mathbb{E}(X_1), \ldots, \mathbb{E}(X_d))^*$$

et sa **matrice de variance-covariance** *var(X) par*

$$var(X) = \mathbb{E}((X - \mathbb{E}(X)) \times (X - \mathbb{E}(X))^*).$$

Notons que $\mathrm{var}(X)$ est symétrique et

$$\forall\, (i,j) \in \{1, \ldots, d\}^2, \quad \mathrm{var}(X)_{ij} = \mathrm{cov}(X_i, X_j).$$

Remarque 2.1. *Si $(X_1, \ldots, X_n)$ est un n-échantillon de loi gausienne alors on a évidemment que $X = (X_1, \ldots, X_n)^*$ est un vecteur gaussien dont la matrice de variance-covariance est proportionnelle à I_d.*

2.1.2 Propriétés des vecteurs gaussiens

Donnons la fonction caractéristique d'un vecteur gaussien et les conséquences importantes qui en découlent.

Théorème 2.2. *Soit $X = (X_1, \ldots, X_d)^*$ un vecteur gaussien. On note $m = \mathbb{E}(X)$ et $\Sigma = var(X)$. On a que X admet pour fonction caractéristique la fonction*

$$\forall\, t \in \mathbb{R}^d, \quad \phi_X(t) = \mathbb{E}[\exp(it^*X)] = \exp\left(it^*m - t^*\Sigma t\right).$$

La loi de X est donc entièrement déterminée par m et Σ. On note $X \sim \mathcal{N}(m, \Sigma)$.

Preuve : On note que

$$\forall\, t \in \mathbb{R}^d, \quad t^*X \sim \mathcal{N}(t^*m, t^*\Sigma t).$$

□

Proposition 2.1. (Propriété de linéarité)
Soit $X = (X_1, \dots, X_d)^*$ un vecteur gaussien. On note $m = \mathbb{E}(X)$ et $\Sigma = var(X)$. On a pour toute matrice A possédant d colonnes et pour tout vecteur $b \in \mathbb{R}^d$,

$$AX + b \sim \mathcal{N}(Am + b, A\Sigma A^*).$$

Preuve : Elle découle du Théorème 2.2 □

Proposition 2.2. (Propriété pour l'indépendance)
Soit $X = (X_1, \dots, X_d)^*$ un vecteur gaussien. On note $m = \mathbb{E}(X)$ et $\Sigma = var(X)$. Pour tout $(i, j) \in \{1, \dots, d\}^2$ tel que $i \neq j$, X_i et X_j sont indépendantes si et seulement si $cov(X_i, X_j) = 0$.

Preuve : Elle découle du Théorème 2.2 □

Remarque 2.2. *Les composantes d'un vecteur gaussien sont des variables aléatoires gaussiennes mais la réciproque est fausse. En effet, on considère $X \sim \mathcal{N}(0, 1)$ et $\varepsilon \sim Ber(0.5)$ indépendante de X. Alors $X_1 = X$ et $X_2 = (2\varepsilon - 1)X$ sont des variables gaussiennes mais $(X_1, X_2)^*$ n'est pas un vecteur gaussien. Notons que $cov(X_1, X_2) = 0$ mais que X_1 et X_2 ne sont pas indépendantes.*

Proposition 2.3. (Propriété pour l'espérance conditionnelle)
Soit $(Y, X_1, \dots, X_d)$ un vecteur gaussien alors $\mathbb{E}(Y|X_1, \dots, X_d)$ est une fonction affine de $(X_1, \dots, X_d)$.

Preuve : Soit $p_{1,X_1,\dots,X_d}(Y)$ la projection de Y sur $\text{vect}(1, X_1, \dots, X_d)$ pour le produit scalaire associé à l'espérance. Donc $\mathbb{E}[(Y - p_{1,X_1,\dots,X_d}(Y))Z] = 0$ pour toute variable $Z \in \text{vect}(1, X_1, \dots, X_d)$. Avec $Z = 1$, on déduit que $\mathbb{E}[Y - p_{1,X_1,\dots,X_d}(Y)] = 0$. Puis, pour toute variable $Z \in \{X_1, \dots, X_d\}$, $(Y - p_{1,X_1,\dots,X_d}(Y), X_1, \dots, X_d)$ étant un vecteur gaussien, $0 = \mathbb{E}[(Y - p_{1,X_1,\dots,X_d}(Y))Z] = \text{cov}(Y - p_{X_1,\dots,X_d}(Y), Z)$ montre que $Y - p_{1,X_1,\dots,X_d}(Y)$ et Z sont indépendantes. Donc $Y - p_{1,X_1,\dots,X_d}(Y)$ est indépendante de toutes fonction de $(X_1, \dots, X_d)$ et $p_{1,X_1,\dots,X_d}(Y) = \mathbb{E}(Y|X_1, \dots, X_d)$. □

A l'aide de la fonction caractéristique, on démontre le TCL vectoriel.

Théorème 2.3. *Soient $X_1, \dots, X_n$ des vecteurs aléatoires de $\mathbb{R}^d$ i.i.d. admettant un moment d'ordre 2. On note m leur espérance et Γ leur matrice de variance-covariance. Alors,*

$$\sqrt{n}(\overline{X}_n - m) \overset{n\to+\infty}{\longrightarrow} \mathcal{N}(0, \Gamma) \text{ en loi.}$$

Preuve : On calcule pour tout n la fonction caractéristique de $Z_n = \sqrt{n}(\overline{X}_n - m)$:

$$\forall\, t \in \mathbb{R}^d, \quad \phi_{Z_n}(t) = \mathbb{E}[\exp(it^* Z_n)].$$

On a par le TCL,

$$t^*Z_n \overset{n\to+\infty}{\longrightarrow} \mathcal{N}(0, t^*\Gamma t) \text{ en loi.}$$

Donc

$$\forall\, t \in \mathbb{R}^d, \quad \phi_{Z_n}(t) \overset{n\to+\infty}{\longrightarrow} \exp\left(-\frac{1}{2}t^*\Gamma t\right).$$

□

Théorème 2.4. *Soit $X = (X_1, \ldots, X_d)^*$ un vecteur gaussien. On note $m = \mathbb{E}(X)$ et $\Sigma = var(X)$. X admet une densité f par rapport à la mesure de Lebesgue sur $\mathbb{R}^d$ si et seulement $det(\Sigma) \neq 0$.*

- *Si $det(\Sigma) = 0$, la loi de $X - m$ est presque sûrement portée par un espace vectoriel engendré par les vecteurs propres associés aux valeurs propres non nulles de Σ.*
- *Si $det(\Sigma) \neq 0$,*

$$\forall\, x \in \mathbb{R}^d, \quad f(x) = \left(\frac{1}{\sqrt{2\pi}}\right)^d \frac{1}{\sqrt{det(\Sigma)}} \exp\left(-\frac{(x-m)^*\Sigma^{-1}(x-m)}{2}\right).$$

Preuve : La matrice Σ est symétrique. Donc il existe U une matrice orthogonale (composée des vecteurs propres de Σ notés $u_1, u_2, \ldots, u_d$) et il existe $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r > 0$ ($r = \text{rang}(\Sigma) \leq d$) tels que

$$\Sigma = U\Gamma U^*,$$

avec

$$\Gamma = \begin{pmatrix} \lambda_1 & & & & & \\ & \ddots & & & & \\ & & \lambda_r & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{pmatrix}.$$

Si $\det(\Sigma) = 0$, on a $r < d$. Pour $i \in \{r+1, \ldots, d\}$, $\mathbb{E}[(u_i^*(X-m))^2] = u_i^*\Sigma u_i = 0$. Donc $u_i^*(X-m) = 0$ p.s. et $X - m$ prend ses valeurs dans $\text{vect}(u_1, \ldots, u_r)$ qui est de mesure de Lebesgue nulle dans $\mathbb{R}^d$.
Si $\det(\Sigma) \neq 0$,, $U\sqrt{\Gamma}$ est inversible. On pose $Y \sim \mathcal{N}(0, I_d)$. Alors $U\sqrt{\Gamma}Y + m \sim X$. Pour toute fonction g continue bornée,

$$\begin{aligned}
\mathbb{E}(g(X)) &= \mathbb{E}(g(U\sqrt{\Gamma}Y + m)) \\
&= \int_{\mathbb{R}^d} g(U\sqrt{\Gamma}y + m) \left(\frac{1}{\sqrt{2\pi}}\right)^d \exp\left(-\frac{\|y\|^2}{2}\right) dy \\
&= \int_{\mathbb{R}^d} g(x) \left(\frac{1}{\sqrt{2\pi}}\right)^d \frac{1}{\sqrt{\det(\Sigma)}} \exp\left(-\frac{(x-m)^*\Sigma^{-1}(x-m)}{2}\right) dx.
\end{aligned}$$

2.2 Théorème de Cochran, lois du χ^2 et de Student

Dans cette section, nous nous placerons dans $\mathbb{R}^d$ muni du produit scalaire euclidien et on notera $\|.\|$ la norme euclidienne dans $\mathbb{R}^d$.

Définition 2.3. *(Cottrell) Soit X un vecteur gaussien de $\mathbb{R}^d$ tel que $\mathbb{E}(X) = m$ et $var(X) = I_d$. La loi de $\|X\|^2$ ne dépend que de d et $\|m\|$. On note*

$$\|X\|^2 \sim \chi^2(d, \|m\|^2)$$

et on dit que $\|X\|^2$ suit une **loi du χ^2 (décentrée si $\|m\| \neq 0$)**. *d est le* **nombre de degrés de liberté**, *$\|m\|^2$ est le* **paramètre de décentrage**.
Lorsque $\|m\| = 0$, on note plus simplement,

$$\|X\|^2 \sim \chi^2(d).$$

Preuve : Soit $Y \in \mathbb{R}^d$ tel que $Y \sim \mathcal{N}(m', I_d)$ avec $\|m\| = \|m'\|$. Il existe U matrice orthogonale telle que $m = Um'$. Donc $UY \sim \mathcal{N}(m, I_d) \sim X$ et

$$\|Y\|^2 = \|UY\|^2 \sim \|X\|^2.$$

□

On a la proposition suivante.

Proposition 2.4. *Si $Z_d \sim \chi^2(d)$, on montre que la densité de Z_d est la fonction f telle que*

$$\forall\, x \in \mathbb{R}, \quad f(x) = \frac{\exp(-x/2)x^{d/2-1}}{2^{d/2}\Gamma(d/2)} 1_{\mathbb{R}_+}(x),$$

avec

$$\forall\, a > 0, \quad \Gamma(a) = \int_0^{+\infty} e^{-x}x^{a-1}dx.$$

On a :

$$\mathbb{E}(Z_d) = d, \quad var(Z_d) = 2d.$$

Preuve : s'obtient par le calcul. □

Enonçons le résultat principal de cette section.

Théorème 2.5. *(Cochran - DDC) Soit $E_1 \oplus \cdots \oplus E_r$ une décomposition de $\mathbb{R}^d$ en sous-espaces deux à deux orthogonaux de dimension respective $d_1, \ldots, d_r$. Si $X \sim \mathcal{N}(m, I_d)$, les vecteurs aléatoires $X_{E_1}, \ldots, X_{E_r}$, projections orthogonales de X sur $E_1, \ldots, E_r$ sont indépendants. Les variables aléatoires $\|X_{E_1}\|^2, \ldots, \|X_{E_r}\|^2$ sont indépendantes et*

$$(\|X_{E_1}\|^2, \ldots, \|X_{E_r}\|^2)^* \sim (\chi^2(d_1, \|m_{E_1}\|^2), \ldots, \chi^2(d_r, \|m_{E_d}\|^2))^*$$

où $m_{E_1}, \ldots, m_{E_r}$, sont les projections de m sur $E_1, \ldots, E_r$.

Preuve : Soit $(e_{j1}, \dots, e_{jd_j})$ une base orthonormée de E_j. On a

$$\forall\, j \in \{1, \dots, d\}, \quad X_{E_j} = \sum_{k=1}^{d_j} e_{jk} e_{jk}^* X.$$

Les variables $e_{jk}^* X$ sont indépendantes de loi $\mathcal{N}(e_{jk}^* m, 1)$ donc les vecteurs aléatoires $X_{E_1}, \dots, X_{E_r}$ sont indépendants. Pour achever la preuve, on remarque que

$$\forall\, j \in \{1, \dots, d\}, \quad \|X_{E_j}\|^2 = \sum_{k=1}^{d_j} (e_{jk}^* X)^2.$$

□

Une application importante du théorème de Cochran est la suivante.

Proposition 2.5. *Soit $X = (X_1, \dots, X_n)$ un n-échantillon de loi $\mathcal{N}(\mu, \sigma^2)$. Reprenons les estimateurs obtenus par la méthode des moments ou par le principe du maximum de vraisemblance pour l'estimation de μ et σ^2 :*

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

Alors, on a :

- *$\overline{X}_n$ et S_n^2 sont des variables aléatoires indépendantes.*
- *Les lois des ces variables sont explicites :*

$$\overline{X}_n \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right), \quad \frac{nS_n}{\sigma^2} \sim \chi^2(n-1).$$

Preuve : On pose pour tout $i \in \{1, \dots, n\}$, $Y_i = \sigma^{-1}(X_i - \mu)$. On a alors que $(Y_1, \dots, Y_n)$ est un n-échantillon de loi $\mathcal{N}(0, 1)$. On pose ensuite $e = (1, \dots, 1)^*$ et $E = \text{vect}(e)$. On a alors

$$\mathbb{R}^n = E \oplus E^\perp.$$

Les projections de $Y = (Y_1, \dots, Y_n)^*$ sur E et $E^\perp$, Y_E et $Y_{E^\perp}$ sont indépendantes et valent

$$Y_E = \frac{1}{n} \sum_{i=1}^{n} Y_i \times e, \quad Y_{E^\perp} = \begin{pmatrix} Y_1 - \frac{1}{n}\sum_{i=1}^n Y_i \\ \vdots \\ Y_n - \frac{1}{n}\sum_{i=1}^n Y_i \end{pmatrix}.$$

On a

$$\frac{1}{\sigma}(\overline{X}_n - \mu) \times e = Y_E, \quad \frac{nS_n}{\sigma^2} = \|Y_{E^\perp}\|^2.$$

□

Ce résultat nous permet de construire des intervalles de confiance pour l'estimation de μ et σ^2 à l'aide de la définition suivante.

Définition 2.4. *Si X et Y sont deux variables aléatoires indépendantes telles que*
- $X \sim \mathcal{N}(\mu, 1)$
- $Y \sim \chi^2(d)$,

alors la loi de la variable

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

est appelée **loi de Student (décentrée si $\mu \neq 0$) à d degrés de liberté**. *On note*

$$Z \sim t(d, \mu).$$

Si **le paramètre de décentrage $\mu = 0$**, *on note plus simplement*

$$Z \sim t(d).$$

Proposition 2.6. *Si $Z_d \sim t(d)$, on montre que la densité de Z_d est la fonction f telle que*

$$\forall\, x \in \mathbb{R}, \quad f(x) = \frac{\Gamma((d+1)/2)}{\sqrt{d\pi}\Gamma(d/2)} \left(1 + \frac{x^2}{d}\right)^{-(d+1)/2},$$

avec

$$\forall\, a > 0, \quad \Gamma(a) = \int_0^{+\infty} e^{-x} x^{a-1} dx.$$

Pour $d > 1$, on a :

$$\mathbb{E}(Z_d) = 0.$$

Pour $d > 2$, on a :

$$var(Z_d) = \frac{d}{d-2}.$$

On a

$$\lim_{d \to +\infty} Z_d = Z \text{ en loi,}$$

où $Z \sim \mathcal{N}(0, 1)$.

Preuve : les premiers points s'obtiennent par le calcul. Pour le dernier point, on peut utiliser le lemme de Slutsky, on peut aussi utiliser la formule de Stirling :

$$\Gamma(x+1) \overset{x \to +\infty}{\sim} \sqrt{2\pi x} \left(\frac{x}{e}\right)^x$$

et le **lemme de Scheffé**.

Lemme 2.1. *Soit $(f_n)_n$ est une suite de densités de probabilité par rapport à la mesure de Lebesgue qui converge simplement vers une densité f. Alors si pour tout n, Z_n est une variable aléatoire de densité f_n et si Z est une variable aléatoire de densité f, on a*

$$\lim_{n \to +\infty} Z_n = Z \text{ en loi.}$$

Preuve du lemme : On pose pour tout n, $g_n = f - f_n$. On a

$$\forall\, x \in \mathbb{R}, \quad 0 \leq g_n(x) 1_{g_n(x)\geq 0} \leq f$$

et par le théorème de convergence dominée

$$\lim_{n\to+\infty} \int g_n(x) 1_{g_n(x)\geq 0} dx = 0.$$

Comme $\int g_n(x)dx = 0$,

$$\begin{aligned} \int |g_n(x)|dx &= \int g_n(x)1_{g_n(x)\geq 0}dx - \int g_n(x)1_{g_n(x)<0}dx \\ &= 2\int g_n(x)1_{g_n(x)\geq 0}dx \\ &\longrightarrow 0 \end{aligned}$$

Donc pour tout $t \in \mathbb{R}$,

$$\lim_{n\to+\infty} \left| \int (f_n(x) - f(x)) 1_{x\leq t} dx \right| \leq \lim_{n\to+\infty} \int |g_n(x)|dx = 0.$$

□

Comme la loi normale, la loi de Student est symétrique mais ses queues sont plus épaisses que celles de la loi normale. On déduit de la définition précédente que

$$\frac{\sigma^{-1}\sqrt{n}(\overline{X}_n - \mu)}{\sigma^{-1}\sqrt{\frac{nS_n}{n-1}}} = \frac{(\overline{X}_n - \mu)}{\sqrt{\frac{S_n}{n-1}}} \sim t(n-1).$$

En notant $t_{n-1,1-\alpha/2}$ le quantile d'ordre $1-\alpha/2$ pour la loi $t(n-1)$ et $c_{n-1,1-\alpha}$ le quantile d'ordre $1-\alpha$ pour la loi $\chi^2(n-1)$, un intervalle de confiance de niveau de confiance exactement égal à $1-\alpha$ pour μ est :

$$I_{n,\alpha} = \left[\overline{X}_n - t_{n-1,1-\alpha/2}\sqrt{\frac{S_n}{n-1}}, \overline{X}_n + t_{n-1,1-\alpha/2}\sqrt{\frac{S_n}{n-1}} \right]$$

et un intervalle de confiance de niveau de confiance exactement égal à $1-\alpha$ pour σ^2 est :

$$J_{n,\alpha} = \left[\frac{nS_n}{c_{n-1,1-\alpha}}, +\infty \right[.$$

On déduit de ces intervalles de confiance les tests de taille α de $\mu = \mu_0$ contre $\mu \neq \mu_0$ et de $\sigma^2 = \sigma_0^2$ contre $\sigma^2 < \sigma_0^2$. Notons que l'on obtient une région de confiance de niveau de confiance $1-2\alpha$ pour l'estimation de $\theta = (\mu, \sigma^2)$ en considérant $I_{n,\alpha} \times J_{n,\alpha}$.

2.3 Test d'ajustement du χ^2

Dans cette section, on considère une variable aléatoire discrète X à valeurs dans $\{a_1, \dots, a_d\}$. On se donne d réels strictement positifs $p_1, \dots, p_d$ tels que $\sum_{i=1}^d p_i = 1$ et on désire tester

$$H_0 : \quad \forall\, i \in \{1, \dots, d\}, \quad \mathbb{P}(X = a_i) = p_i$$

contre

$$H_1 : \quad \exists\, i \in \{1, \dots, d\}, \quad \mathbb{P}(X = a_i) \neq p_i.$$

Pour cela, on dispose d'un n-échantillon $(X_1, \dots, X_n)$ de même loi que X. On utilise la méthodes des moments pour estimer p_i et on note

$$\forall\, i \in \{1, \dots, d\}, \quad N_{ni} = \sum_{j=1}^n 1_{X_j = a_i}, \quad \hat{p}_i = \frac{N_{ni}}{n}.$$

Sous H_0, pour tout $i \in \{1, \dots, d\}$, $\hat{p}_i$ est un estimateur fortement consistant et sans biais de p_i. Donc si H_0 est vraie, il y a tout lieu de penser que $\hat{p} = (\hat{p}_1, \dots, \hat{p}_d)^*$ sera "proche" de $p = (p_1, \dots, p_d)^*$. Comment mesurer la distance entre $\hat{p}$ et p ? On introduit la **pseudo-distance du** χ^2 entre $\hat{p}$ et p :

$$D_n^2(\hat{p}, p) = n \sum_{i=1}^d \frac{(\hat{p}_i - p_i)^2}{p_i}.$$

Lorsque n est grand, sa limite est connue et surtout indépendante de p, ce qui va nous permettre de résoudre notre problème de test. On a en effet le théorème suivant :

Théorème 2.6. *On a :*

- *sous* H_0 *:*

$$D_n^2(\hat{p}, p) \overset{n \to +\infty}{\longrightarrow} \chi^2(d-1) \text{ en loi,}$$

- *sous* H_1 *:*

$$D_n^2(\hat{p}, p) \overset{n \to +\infty}{\longrightarrow} +\infty \text{ p.s.}$$

Preuve : Remarque inutile : La loi du vecteur $N_n = (N_{n1}, \dots, N_{nd})^*$ est la loi multinomiale $\mathcal{M}(n, p)$:

$$\forall\, \tilde{n} = (n_1, \dots, n_d)^* \in \mathbb{N}^d \text{ avec } \sum_{j=1}^d n_j = n, \quad \mathbb{P}(N_n = \tilde{n}) = \frac{n!}{n_1! \cdots n_d!} p_1^{n_1} \cdots p_d^{n_d}.$$

On pose

$$\forall\, j \in \{1, \dots, n\}, \quad Z_j = \left(\frac{1}{\sqrt{p_1}}(1_{X_j = a_1} - p_1), \dots, \frac{1}{\sqrt{p_d}}(1_{X_j = a_d} - p_d) \right)^*.$$

Par le TCL vectoriel,

$$\frac{1}{\sqrt{n}}\left(\sum_{j=1}^{n} Z_j\right) \overset{n\to+\infty}{\longrightarrow} \mathcal{N}(0, I_d - \sqrt{p}\sqrt{p}^*) \text{ en loi,}$$

avec $\sqrt{p} = (\sqrt{p_1}, \ldots, \sqrt{p_d})^*$. Donc

$$\sqrt{n}\left(\frac{1}{\sqrt{p_1}}(\hat{p}_1 - p_1), \ldots, \frac{1}{\sqrt{p_d}}(\hat{p}_1 - p_d)\right)^* \overset{n\to+\infty}{\longrightarrow} \mathcal{N}(0, I_d - \sqrt{p}\sqrt{p}^*) \text{ en loi.}$$

En utilisant la fonction continue f définie par

$$\forall\, x = (x_1, \ldots, x_d)^*, \quad f(x) = \|x\|^2 = \sum_{j=1}^{d} x_j^2,$$

on obtient

$$D_n^2(\hat{p}, p) \overset{n\to+\infty}{\longrightarrow} f(V) \text{ en loi,}$$

où V est une variable aléatoire telle que $V \sim \mathcal{N}(0, I_d - \sqrt{p}\sqrt{p}^*)$ et qui a donc même loi que la projection de $W \sim \mathcal{N}(0, I_d)$ sur $(\text{vect}(\sqrt{p}))^\perp$. Donc

$$f(V) \sim \chi^2(d-1).$$

□

Pour tester H_0 contre H_1, on considère donc le test asymptotique de taille $1-\alpha$

$$\phi(X_1, \ldots, X_n) = 1_{D_n^2(\hat{p},p) > c_{d-1,1-\alpha}}$$

où $c_{d-1,1-\alpha}$ est le quantile d'ordre $1-\alpha$ de la loi $\chi^2(d-1)$. Notons que la puissance du test tend vers 1 quand n tend vers $+\infty$.

Remarque 2.3. *L'approximation par la loi limite est correcte si pour tout $i \in \{1, \ldots, d\}$, $np_i \geq 5$. Si ce n'est pas le cas, il faut effectuer un regroupement par classes.*

Remarque 2.4. *On peut utiliser ce test lorsque la loi de X est continue. Si X est à valeurs dans Ω, on construit une partition finie de Ω et on applique ce qui précède. Tout le problème porte sur le choix de cette partition.*

Exemple historique : (Wasserman) Pour tester sa théorie génétique, Mendel croisa des pois tous jaunes et lisses et obtint à la première génération des pois jaunes ou verts et lisses ou ridés. Plus précisément, il obtint 315 pois jaunes et lisses, 108 pois verts et lisses, 101 pois jaunes et ridés et 32 pois verts et ridés. Est ce que ces observations confirment ou infirment la théorie mendélienne ? Sous cette approche, la proportion p de chacune des 4 classes précédentes est $p = (\frac{9}{16}, \frac{3}{16}, \frac{3}{16}, \frac{1}{16})^*$. On teste donc

$$H_0 : p = \left(\frac{9}{16}, \frac{3}{16}, \frac{3}{16}, \frac{1}{16}\right)^*$$

contre

$$H_1 : p \neq \left(\frac{9}{16}, \frac{3}{16}, \frac{3}{16}, \frac{1}{16}\right)^* .$$

On a $c_{3,0.95} = 7.815$. Comme sous H_0, $D^2_{556}(\hat{p}, p) = 0.47$, on accepte H_0. La p-valeur pour ce test est en fait de 0.93.

Des extensions des problèmes d'estimation et de tests qui utilisent les propriétés des vecteurs gaussiens sont envisagés dans le chapitre suivant.

Chapitre 3

Modèle linéaire

Le modèle linéaire est un cadre extrêmement simple qui permet d'appliquer de manière naturelle les notions de statistique vues dans le premier chapitre (estimation, test,...). Les techniques de démonstration s'appuient essentiellement sur les outils de l'algèbre linéaire. L'utilisation des lois de Student, du $\chi 2$ et de Fisher vous nous permettre de nous affranchir de la contrainte de l'asymptotique qui sera donc très peu présente dans ce chapitre. Des exemples concrets illustrant les résultats théoriques vus dans ce chapitre seront étudiés en TP.

3.1 Généralités

Dans toute la suite, on se placera dans $\mathbb{R}^n$ avec $n \in \mathbb{N}^*$ qui sera la taille du vecteur d'observations et on considérera le produit scalaire euclidien standard de $\mathbb{R}^n$. On notera $\|.\|$ la norme euclidienne associée.

3.1.1 Définitions

Définition 3.1. *Soient $n \in \mathbb{N}^*$ et Y un vecteur d'observations de $\mathbb{R}^n$. On dit alors que Y suit un modèle linéaire gaussien si et seulement si*

$$Y = m + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I_n), \tag{3.1}$$

où $m \in \mathbb{R}^n$ et $\sigma > 0$ sont inconnus mais avec $m \in V$ où V est un sous-espace vectoriel connu de $\mathbb{R}^n$ de dimension $p \in \mathbb{N}^*$.

En toutes généralités, V peut être égal à $\mathbb{R}^n$, mais ce qui va nous intéresser dans ce chapitre c'est le cas où $p << n$. En notant $\theta = (m, \sigma^2)$ et $\mathbb{P}_\theta = \mathcal{N}(m, \sigma^2 I_n)$, le modèle statistique associé à ce problème est donc $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), \mathbb{P}_\theta, \theta \in V \times \mathbb{R}_+^*)$. On adoptera le plus souvent une écriture matricielle en considérant V comme étant l'image de p vecteurs colonnes $X_1, \ldots, X_p$. Le modèle est alors formulé de la façon équivalente suivante :

$$Y = X\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I_n), \tag{3.2}$$

avec $X = (X_1, \ldots, X_p)$ et $\beta \in \mathbb{R}^p$. Dans ce cas là, en notant $\theta = (\beta, \sigma^2)$ et $\mathbb{P}_\theta = \mathcal{N}(X\beta, \sigma^2 I_n)$, le modèle statistique est $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), \mathbb{P}_\theta, \theta \in \mathbb{R}^p \times \mathbb{R}_+^*)$. L'écriture (3.2) sera fréquemment utilisée. Il est en effet plus facile d'interpréter les paramètres concrètement, comme l'illustrent les exemples suivants.

Exemple : Considérons le rendement d'une réaction chimique qui dépend linéairement de la température et du pH. On répète la réaction n fois de manière indépendante. On note Y_i le rendement de la i-ème réaction et Z_i^1 et Z_i^2 les $i-$ème coordonnées des vecteurs pH et température. On écrit le **modèle de régression linéaire multiple** suivant :

$$\forall\, i \in \{1, \ldots, n\}, \quad Y_i = \mu + \beta_1 Z_i^1 + \beta_2 Z_i^2 + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2).$$

En notant $Y = (Y_1, \ldots, Y_n)^*$, $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^*$, $\beta = (\mu, \beta_1, \beta_2)^*$ et

$$X = \begin{pmatrix} 1 & Z_1^1 & Z_1^2 \\ \vdots & \vdots & \vdots \\ 1 & Z_n^1 & Z_n^2 \end{pmatrix}$$

on obtient bien la modélisation (3.2) et dans ce cas là, $p = 3$.

Exemple : Considérons la variable du rendement associé à la culture du blé. On note μ le rendement moyen. On peut **expliquer** cette variable par deux **facteurs** : la variété et l'exposition. On suppose que ces facteurs ont chacun deux niveaux et pour chacun de ces niveaux, on observe le rendement sur q parcelles que l'on suppose indépendantes. On écrit le **modèle de l'analyse de la variance** suivant :

$$\begin{array}{rrcll}
\forall\, i \in \{1, \ldots, q\}, & Y_i & = & \mu + a_1 + b_1 + \varepsilon_i, & \varepsilon_i \sim \mathcal{N}(0, \sigma^2) \\
\forall\, i \in \{1, \ldots, q\}, & Y_{i+q} & = & \mu + a_1 + b_2 + \varepsilon_{i+q}, & \varepsilon_{i+q} \sim \mathcal{N}(0, \sigma^2) \\
\forall\, i \in \{1, \ldots, q\}, & Y_{i+2q} & = & \mu + a_2 + b_1 + \varepsilon_{i+2q}, & \varepsilon_{i+2q} \sim \mathcal{N}(0, \sigma^2) \\
\forall\, i \in \{1, \ldots, q\}, & Y_{i+3q} & = & \mu + a_2 + b_2 + \varepsilon_{i+3q}, & \varepsilon_{i+3q} \sim \mathcal{N}(0, \sigma^2).
\end{array}$$

En notant $n = 4q$, $Y = (Y_1, \dots, Y_n)^*$, $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^*$, $\beta = (\mu, a_1, a_2, b_1, b_2)^*$ et

$$X = \begin{pmatrix} 1 & 1 & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 0 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & 0 & 1 \end{pmatrix}$$

on obtient bien la modélisation (3.2) et dans ce cas là, $p = 5$. Bien entendu, la modélisation et les hypothèses qui sous-tendent le modèle doivent être analysées et discutées en particulier à l'aide de représentations graphiques (cf Section 3.5).

Il faut noter que dans le cadre de l'écriture (3.2), le modèle est identifiable si et seulement si X est injective (ce qui équivaut à $\text{rang}(X) = p$ quand $p \leq n$). Si X n'est pas injective, on pose des contraintes d'identifiabilité pour rendre le modèle identifiable. On propose une contrainte de la forme $\beta \in K$ où K est une sous-espace vectoriel de $\mathbb{R}^p$ et tel que pour tout $m \in \text{Im}(X)$ il existe un unique $\beta \in K$ tel que $m = X\beta$. On peut prendre par exemple $K = \text{Ker}(X)^\perp$. Dans l'exemple de la régression, le modèle est identifiable si Z^1 et Z^2 ne sont pas colinéaires et s'il existe i, i', j et j' tels que $Z^1_i \neq Z^1_{i'}$ et $Z^2_j \neq Z^2_{j'}$. Dans l'exemple précédent de l'analyse de la variance, le modèle n'est pas identifiable car il est surparamétré. La contrainte d'identifiabilité que l'on peut choisir est $\beta \in K = \{(\mu, a_1, a_2, b_1, b_2)^* : \quad a_1 + a_2 + b_1 + b_2 = 0\}$. **Jusqu'à la fin de ce chapitre, on supposera le modèle identifiable et X injective**.

3.1.2 Estimation

Sous le modèle (3.1), on a le résultat suivant.

Théorème 3.1. *Sous le modèle linéaire gaussien*

$$Y = m + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I_n),$$

où $\sigma > 0$, $m \in V$ où V est un sous-espace vectoriel connu de $\mathbb{R}^n$ de dimension $p \in \mathbb{N}^*$, on a

- *l'estimateur du maximum de vraisemblance de* $\theta = (m, \sigma^2)$ *est donné par* $\hat{\theta} = (\hat{m}, s_n^2)$ *avec*
$$\hat{m} = \Pi_V Y, \quad s_n^2 = \frac{1}{n}\|Y - \Pi_V Y\|^2,$$
où Π_V *est la matrice de projection sur* V*.*
- *Ces estimateurs sont indépendants. On a*
$$\hat{m} \sim \mathcal{N}(m, \sigma^2 \Pi_V), \quad \frac{n s_n^2}{\sigma^2} \sim \chi^2(n - dim(V))$$
et $\hat{m}$ *est sans biais, mais* s_n^2 *est biaisé et asymptotiquement sans biais.*
- *Un estimateur sans biais de* σ^2 *est donné par*
$$\hat{\sigma}^2 = \frac{1}{n - dim(V)}\|Y - \Pi_V Y\|^2.$$

Preuve : La vraisemblance s'écrit :

$$\forall\, \theta \in V \times \mathbb{R}_+^*, \quad V_Y(\theta) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\|Y - m\|^2\right).$$

Par ailleurs, par Pythagore, pour $m \in V$,

$$\|Y - m\|^2 = \|Y - \Pi_V Y\|^2 + \|\Pi_V Y - m\|^2$$

et

$$\hat{m} = \Pi_V Y = m + \Pi_V \varepsilon.$$

On obtient l'estimateur du maximum de vraisemblance de σ^2 en dérivant. Comme $I_n - \Pi_V$ est le projecteur orthogonal sur $\mathrm{Ker}(\Pi_V) = \mathrm{Im}(\Pi_V)^\perp$, le théorème de Cochran permet d'achever la démonstration. □

Il est important de noter que $\hat{m}$ vérifie sans hypothèse sur la loi des erreurs

$$\hat{m} = \arg\min_{m \in V} \|Y - m\|.$$

Il est donc **l'estimateur des moindres carrés ordinaires**. A présent donnons les résultats d'estimation dans le cadre du modèle (3.2).

Théorème 3.2. *Sous le modèle linéaire gaussien*

$$Y = X\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I_n),$$

avec $X \in M_{np}(\mathbb{R})$ *injective* ($p \leq n$)*,* $\sigma > 0$ *et* $\beta \in \mathbb{R}^p$*, on a :*

- *l'estimateur du maximum de vraisemblance de* $\theta = (\beta, \sigma^2)$ *est donné par* $\hat{\theta} = (\hat{\beta}, s_n^2)$ *avec*
$$\hat{\beta} = (X^* X)^{-1} X^* Y, \quad s_n^2 = \frac{1}{n}\|Y - X\hat{\beta}\|^2.$$

– *Ces estimateurs sont indépendants. On a*

$$\hat{\beta} \sim \mathcal{N}\beta, \sigma^2(X^*X)^{-1}), \quad \frac{ns_n^2}{\sigma^2} \sim \chi^2(n-p)$$

et $\hat{\beta}$ est sans biais, mais s_n^2 est biaisé et asymptotiquement sans biais.

– *Un estimateur sans biais de σ^2 est donné par*

$$\hat{\sigma}^2 = \frac{1}{n-p}\|Y - X\hat{\beta}\|^2.$$

– *Parmi les estimateurs linéaires et sans biais de β, $\hat{\beta}$ est celui dont la matrice de variance-covariance est "minimale" pour la relation d'ordre naturelle sur les matrices symétriques (propriété de Gauss-Markov).*

Preuve : Pour les trois premiers points, on utilise le théorème précédent. Il suffit alors d'observer que si $V = \mathrm{Im}(X)$,

$$\Pi_V = X(X^*X)^{-1}X^*$$

et d'utiliser l'injectivité de X (qui équivaut à l'inversibilité de X^*X). Pour le dernier point, notons $\tilde{\beta} = CY$ un estimateur linéaire sans biais de β. On a pour tout $\beta \in \mathbb{R}^p$, $\mathbb{E}(CY) = CX\beta = \beta$ donc $CX = I_p$, puis

$$\begin{aligned}
\mathrm{var}(\tilde{\beta}) &= \mathrm{var}(CY) \\
&= \sigma^2 CC^* \\
&= \sigma^2(C - (X^*X)^{-1}X^* + (X^*X)^{-1}X^*)(C - (X^*X)^{-1}X^* + (X^*X)^{-1}X^*)^* \\
&= \sigma^2(C - (X^*X)^{-1}X^*)(C - (X^*X)^{-1}X^*)^* + \sigma^2(XX^*)^{-1}.
\end{aligned}$$

□

Ce dernier résultat montre que l'intervalle de confiance pour β que l'on privilégiera naturellement sera celui construit autour de $\hat{\beta}$.

3.2 Régions de confiance et tests fondamentaux

Dans cette section, on se place alternativement dans les modèles

$$Y = X\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I_n),$$

avec $X = (X_1, \ldots, X_p)$ et $\beta \in \mathbb{R}^p$, $p < n$ ou

$$Y = m + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I_n),$$

avec $m \in V$ où V est un sous-espace vectoriel de $\mathbb{R}^n$ de dimension $p < n$. Ce qui est énoncé dans le cadre d'un modèle pourra à chaque fois être transposé dans le cadre de l'autre modèle.

3.2.1 Tests pour la variance

On se donne $\sigma_0 \in \mathbb{R}_+^*$ et une fois n'est pas coutume, on propose un test unilatéral en testant l'hypothèse $H_0 : \sigma = \sigma_0$ contre $H_1 : \sigma > \sigma_0$. On utilise le fait que

$$\frac{(n-p)\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-p).$$

Pour $\alpha \in [0,1]$, on note $c_{n-p,1-\alpha}$ le quantile d'ordre $1-\alpha$ de la loi du χ^2 à $n-p$ degrés de liberté. Le test

$$\phi(Y) = 1_{\hat{\sigma}^2 > \sigma_0^2 (n-p)^{-1} c_{n-p,1-\alpha}}$$

est un test de taille α de H_0 contre H_1. L'intervalle de confiance unilatère de niveau exactement $1-\alpha$ pour l'estimation de σ^2 est donnée par

$$\left[\frac{(n-p)\hat{\sigma}^2}{c_{n-p,1-\alpha}}, +\infty \right[.$$

3.2.2 Test de Student

On se place dans le modèle

$$Y = X\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I_n),$$

avec $X = (X_1, \ldots, X_p)$ et $\beta \in \mathbb{R}^p$, $p < n$. On rappelle la définition suivante.

Définition 3.2. *Si U et V sont deux variables aléatoires indépendantes telles que*
- *$U \sim \mathcal{N}(\mu, 1)$*
- *$V \sim \chi^2(d)$,*

alors la loi de la variable

$$Z = \frac{U}{\sqrt{\frac{V}{d}}}$$

est appelée **loi de Student (décentrée si $\mu \neq 0$) à d degrés de liberté**. *On note*

$$Z \sim t(d, \mu).$$

Si **le paramètre de décentrage $\mu = 0$**, *on note plus simplement*

$$Z \sim t(d).$$

On se donne $c \in \mathbb{R}^p$, $a \in \mathbb{R}$ et on désire tester l'hypothèse $H_0 : c^*\beta = a$ contre $H_1 : c^*\beta \neq a$. Rentre notamment dans ce cadre le test de $\beta_k = 0$ contre $\beta_k \neq 0$ pour $k \in \{1, \ldots, p\}$. On utilise naturellement la statistique

$$T = \frac{c^*\hat{\beta} - a}{\hat{\sigma}\sqrt{c^*(X^*X)^{-1}c}}.$$

On a

$$c^*\hat{\beta} - a \sim \mathcal{N}(c^*\beta - a, \sigma^2 c^*(X^*X)^{-1}c), \quad \frac{(n-p)\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-p)$$

donc comme $\hat{\beta}$ et $\hat{\sigma}$ sont indépendants, $T \sim t(n-p, c^*\beta - a)$ et $T \sim t(n-p)$ sous H_0. Pour $\alpha \in [0,1]$, on note $t_{n-p,1-\alpha/2}$ le quantile d'ordre $1-\alpha/2$ de la loi de Student à $n-p$ degrés de liberté. Le test

$$\phi(Y) = 1_{|T|>t_{n-p,1-\alpha/2}}$$

est un test de taille α de H_0 contre H_1. L'intervalle de confiance bilatère de niveau exactement $1-\alpha$ pour l'estimation de $c^*\beta$ est donnée par

$$\left[c^*\hat{\beta} - t_{n-p,1-\alpha/2}\hat{\sigma}\sqrt{c^*(X^*X)^{-1}c}, c^*\hat{\beta} + t_{n-p,1-\alpha/2}\hat{\sigma}\sqrt{c^*(X^*X)^{-1}c}\right].$$

3.2.3 Test de Fisher d'un sous-modèle

On se place dans le modèle (3.1)

$$Y = m + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I_n),$$

avec $m \in V$ où V est un sous-espace vectoriel de $\mathbb{R}^n$ de dimension $p < n$. On se donne W un sous espace vectoriel de V de dimension $q < p$ et on se propose de tester $H_0 : m \in W$ contre $H_1 : m \in V \setminus W$. Introduisons la loi de Fisher.

Définition 3.3. *Si U et V sont deux variables aléatoires indépendantes telles que*
- *$U \sim \chi^2(p)$, $p \in \mathbb{N}^*$,*
- *$V \sim \chi^2(q)$, $q \in \mathbb{N}^*$,*

alors la loi de la variable

$$Z = \frac{U/p}{V/q}$$

est appelée **loi de Fisher à p et q degrés de liberté**. *On note*

$$Z \sim \mathcal{F}(p,q).$$

On a le résultat suivant.

Théorème 3.3. *Dans le cadre du modèle (3.1), si V et W sont deux sous-espaces vectoriels de $\mathbb{R}^n$ de dimensions respectives p et q tels que $W \subset V$ et $q < p < n$, alors si $m \in W$,*

$$F = \frac{\left(\|Y - \Pi_W Y\|^2 - \|Y - \Pi_V Y\|^2\right)/(p-q)}{\|Y - \Pi_V Y\|^2/(n-p)} = \frac{\left(\|\Pi_W Y - \Pi_V Y\|^2\right)/(p-q)}{(\|Y - \Pi_V Y\|^2)/(n-p)} \sim \mathcal{F}(p-q, n-p),$$

où Π_V et Π_W sont les matrices de projection sur V et W respectivement. De plus, $\Pi_W Y$ est indépendante de F.

Preuve : Elle découle du théorème de Cochran et du théorème de Pythagore. □

Pour $\alpha \in [0,1]$, on note $f_{p-q,n-p,1-\alpha}$ le quantile d'ordre $1-\alpha$ de la loi de Fisher à $p-q$ et $n-p$ degrés de liberté. Le test

$$\phi(Y) = 1_{F > f_{p-q,n-p,1-\alpha}}$$

est un test de taille α de H_0 contre H_1. Il faut noter que si $q = p-1$, alors W est un hyperplan de V et il existe $c \in \mathbb{R}^p$ tel que $m \in W \Leftrightarrow c^* m = 0$. Nous sommes alors amenés à un test de Student étudié dans la section précédente. Il faut noter que ces tests sont équivalents puisque pour $p \in \mathbb{N}^*$, si $Z \sim t(p)$, alors $Z^2 \sim \mathcal{F}(1,p)$.

3.2.4 Test de Wald

On se place dans le modèle (3.2)

$$Y = X\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I_n),$$

avec $X = (X_1, \ldots, X_p)$ et $\beta \in \mathbb{R}^p$, $p < n$. On veut tester dans cette section une hypothèse affine. Pour cela, on se donne une matrice $C \in M_{k,p}(\mathbb{R})$ avec $k \leq p$. On supposera que C^* est injective et donc de rang k. Pour $a \in \mathbb{R}^k$, on désire tester $H_0 : C\beta = a$ contre $H_1 : C\beta \neq a$. Quand $a = 0$, on peut se ramener au test de Fisher d'un sous-modèle, mais ce sous-modèle n'es pas facile à expliciter. Si on veut tester par exemple deux hypothèses linéaires simultanées $\beta_2 = 2\beta_1$ et $\beta_3 = 0$, on prendra

$$C = \begin{pmatrix} -2 & 1 & 0 & 0 & \cdots \\ 0 & 0 & 1 & 0 & \cdots \end{pmatrix}, \quad a = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Théorème 3.4. *Dans le cadre du modèle (3.2), avec les notations précédentes, sous H_0,*

$$W = \frac{\left((C\hat{\beta} - a)^* (C(X^*X)^{-1}C^*)^{-1}(C\hat{\beta} - a)\right)/k}{\|Y - X\hat{\beta}\|^2/(n-p)} \sim \mathcal{F}(k, n-p).$$

Preuve : On a

$$C\hat{\beta} - a \sim \mathcal{N}(C\beta - a, \sigma^2 C(X^*X)^{-1}C^*)$$

et donc sous H_0

$$C\hat{\beta} - a \sim \mathcal{N}(0, \sigma^2 C(X^*X)^{-1}C^*).$$

En observant que $C(X^*X)^{-1}C^* \in M_{k,k}(\mathbb{R})$ est symétrique définie positive (car C^* est injective), on pose

$$\Delta = \sigma^{-1}\sqrt{(C(X^*X)^{-1}C^*)^{-1}}$$

et on calcule sous H_0 :

$$\sigma^{-2}(C\hat{\beta} - a)^*(C(X^*X)^{-1}C^*)^{-1}(C\hat{\beta} - a) = (C\hat{\beta} - a)^*\Delta^2(C\hat{\beta} - a) = \|\Delta(C\hat{\beta} - a)\|_k^2 \sim \chi^2(k)$$

où $\|\cdot\|$ est la norme euclidienne de $\mathbb{R}^k$. On conclut en rappelant que $\hat{\beta}$ est indépendante de $\|Y - X\hat{\beta}\|^2$. □

Pour $\alpha \in [0,1]$, on note $f_{k,n-p,1-\alpha}$ le quantile d'ordre $1-\alpha$ de la loi de Fisher à k et $n-p$ degrés de liberté. Le test

$$\phi(Y) = 1_{W > f_{k,n-p,1-\alpha}}$$

est un test de taille α de H_0 contre H_1. On déduit également un ellipsoïde de confiance de niveau exactement $1-\alpha$ pour l'estimation de $C\beta$:

$$\mathcal{E} = \left\{ a : \quad \frac{\left((C\hat{\beta} - a)^*(C(X^*X)^{-1}C^*)^{-1}(C\hat{\beta} - a)\right)/k}{\|Y - X\hat{\beta}\|^2/(n-p)} \leq f_{k,n-p,1-\alpha} \right\}.$$

3.3 Applications à la régression linéaire

Pour $q \in \mathbb{N}^*$, le **modèle de régression linéaire** s'écrit sous la forme suivante :

$$\forall\, i \in \{1,\dots,n\}, \quad Y_i = \mu + \beta_1 Z_i^1 + \beta_2 Z_i^2 + \cdots + \beta_q Z_i^q + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0,\sigma^2)$$

où les variables ε_i sont indépendantes. Si on pose

$$\forall\, j \in \{1,\dots,q\}, \quad Z^j = (Z_1^j,\dots,Z_n^j)^*,$$

les vecteurs $Z^1,\dots,Z^q$ sont appelés **variables explicatives ou régresseurs**. En notant $Y = (Y_1,\dots,Y_n)^*$, $\varepsilon = (\varepsilon_1,\dots,\varepsilon_n)^*$, $\beta = (\mu,\beta_1,\beta_2,\dots,\beta_q)^*$ et

$$X = \begin{pmatrix} 1 & Z_1^1 & Z_1^2 & \dots & Z_1^q \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & Z_n^1 & Z_n^2 & \dots & Z_n^q \end{pmatrix}$$

on obtient bien la modélisation (3.2) :

$$Y = X\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0,\sigma^2 I_n).$$

Quand $q = 1$, on parle de **régression linéaire simple**. Quand $q > 1$, on parle de **régression linéaire simple**.

Remarque 3.1. *Les vecteurs $Z^1,\dots,Z^q$ peuvent avoir un lien entre elles. Par exemple, chaque coordonnée de chaque vecteur Z_j peut être la j-ème puissance de la coordonnée correspondante d'un vecteur Z. On cherche ainsi un ajustement polynomial de la variable à expliquer. Ce travail peut être exploité pour effectuer de la prédiction.*

On peut dans ce cadre tester l'utilité de régresseurs en appliquant les résultats de la Section 3.2.3. Quitte à intervertir les vecteurs Z^j, on va tester l'hypothèse "les régresseurs $Z^{q'+1},\dots,Z^q$ sont inutiles" contre "au moins un parmi les régresseurs $Z^{q'+1},\dots,Z^q$ est utile" avec $0 \leq q' < q$. Cela revient à tester H_0 : "$\beta_{q'+1}=0,\dots,\beta_q=0$" contre H_1 : il existe $j \in \{q'+1,\dots,q\}$ tel que $\beta_j \neq 0$". Appliquant le Théorème 3.3, avec

$$X_0 = \begin{pmatrix} 1 & Z_1^1 & Z_1^2 & \dots & Z_1^{q'} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & Z_n^1 & Z_n^2 & \dots & Z_n^{q'} \end{pmatrix},$$

sous H_0,

$$F = \frac{\left(\|X_0(X_0^*X_0)^{-1}X_0^*Y - X(X^*X)^{-1}X^*Y\|^2\right)/(q-q')}{(\|Y - X(X^*X)^{-1}X^*Y\|^2)/(n-q-1)} \sim \mathcal{F}(q-q', n-q-1).$$

Pour $\alpha \in [0,1]$, on note $f_{q-q',n-q-1,1-\alpha}$ le quantile d'ordre $1-\alpha$ de la loi de Fisher à $q-q'$ et $n-q-1$ degrés de liberté. Le test

$$\phi(Y) = 1_{F > f_{q-q',n-q-1,1-\alpha}}$$

est un test de taille α de H_0 contre H_1.

Toutes les formules dérivées dans les sections précédentes s'appliquent. En particulier dans le cas de le cas de la **régression simple où** $q=1$, on obtient l'estimateur des moindres carrés de $\beta = (\mu, \beta_1)^*$, $\hat{\beta} = (\hat{\mu}, \hat{\beta}_1)^*$ avec

$$\hat{\beta}_1 = \frac{\mathrm{cov}(Z^1, Y)}{\mathrm{var}(Z^1)} = \frac{\frac{1}{n}\sum_{i=1}^n (Z_i^1 - \bar{Z}^1)(Y_i - \bar{Y})}{\frac{1}{n}\sum_{i=1}^n (Z_i^1 - \bar{Z}^1)^2}, \quad \hat{\mu} = \bar{Y} - \hat{\beta}_1 \bar{Z}^1$$

et

$$\bar{Z}^1 = \frac{1}{n}\sum_{i=1}^n Z_i^1, \quad \bar{Y} = \frac{1}{n}\sum_{i=1}^n Y_i.$$

On a de plus,

$$\hat{\beta} \sim \mathcal{N}\left(\begin{pmatrix} \mu \\ \beta_1 \end{pmatrix}, \frac{\sigma^2}{n\sum_{i=1}^n (Z_i^1 - \bar{Z}^1)^2} \begin{pmatrix} \sum_{i=1}^n (Z_i^1)^2 & -\sum_{i=1}^n Z_i^1 \\ -\sum_{i=1}^n Z_i^1 & n \end{pmatrix} \right).$$

Pour l'estimation sans biais de la variance σ^2, on obtient :

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^n (Y_i - \hat{\mu} - \hat{\beta}_1 Z_i^1)^2.$$

On a

$$\frac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-2).$$

Pour $\alpha \in [0,1]$, le test de taille α de H_0 : "$\beta_1 = 0$" contre H_1 :"$\beta_1 \neq 0$" est donné par

$$\phi(Y) = 1_{F > f_{1,n-2,1-\alpha}}$$

où

$$F = \frac{\sum_{i=1}^n \left(\hat{\mu} + \hat{\beta}_1 Z_i^1 - \bar{Y}\right)^2}{\left(\sum_{i=1}^n \left(Y_i - \hat{\mu} - \hat{\beta}_1 Z_i^1\right)^2\right)/(n-2)} \sim \mathcal{F}(1, n-2)$$

et $f_{1,n-2,1-\alpha}$ est le quantile d'ordre $1-\alpha$ de la loi de Fisher à 1 et $n-2$ degrés de liberté (on applique ce qui précède avec $q = 1$ et $q' = 0$). Remarquons qu'un test strictement équivalent aurait pu être construit à l'aide d'une variable de Student à $n-2$ degrés de liberté.

3.4 Applications à l'analyse de la variance

3.4.1 Analyse de la variance à un facteur

Pour $p \in \mathbb{N}^*$, le **modèle d'analyse de la variance à un facteur** s'écrit sous la forme :

$$\forall\, i \in \{1, \ldots, p\}, \quad \forall\, k \in \{1, \ldots, n_i\}, \quad Y_{ik} = \mu_i + \varepsilon_{ik}, \quad \varepsilon_{ik} \sim \mathcal{N}(0, \sigma^2)$$

où les variables ε_{ik} sont indépendantes et pour tout $i \in \{1, \ldots, p\}$, $n_i \in \mathbb{N}^*$. Par exemple, on explique le rendement associé à la culture du blé en fonction d'un seul **facteur** : la variété. On fait pousser sur n_i parcelles la variété numéro i et on observe le rendement associé. C'est le vecteur $Y_i = (Y_{i1}, \ldots, Y_{in_i})^*$. En notant $n = \sum_{i=1}^p n_i$,

$$Y = (Y_{11}, \ldots, Y_{1n_1}, Y_{21}, \ldots, Y_{2n_2}, \ldots, Y_{p1}, \ldots, Y_{pn_p})^*,$$

$$\varepsilon = (\varepsilon_{11}, \ldots, \varepsilon_{1n_1}, \varepsilon_{21}, \ldots, \varepsilon_{2n_2}, \ldots, \varepsilon_{p1}, \ldots, \varepsilon_{pn_p})^*,$$

$$\beta = (\mu_1, \ldots, \mu_p)^*$$

et

$$X = (X_1, \ldots, X_p),$$

où pour tout $i \in \{1, \ldots, p\}$, X_i est un vecteur dont toutes les coordonnées sont nulles sauf n_i coordonnées valant 1, on obtient bien la modélisation (3.2) :

$$Y = X\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I_n).$$

Dans la suite, on va noter (notation classique en analyse de la variance)

$$\forall\, i \in \{1, \ldots, p\}, \quad Y_{i\cdot} = \frac{1}{n_i} \sum_{k=1}^{n_i} Y_{ik}, \text{ et } Y_{\cdot\cdot} = \frac{1}{n} \sum_{i=1}^{p} \sum_{k=1}^{n_i} Y_{ik}.$$

On obtient l'estimateur des moindres carrés de $\beta = (\mu_1, \dots, \mu_p)^*$, $\hat{\beta} = (\hat{\mu}_1, \dots, \hat{\mu}_p)^*$ avec

$$\hat{\beta} = (Y_{1.}, \dots, Y_{p.})^* .$$

On a de plus,

$$\hat{\beta} \sim \mathcal{N}\left(\beta, \sigma^2 \begin{pmatrix} n_1^{-1} & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & 0 & n_p^{-1} \end{pmatrix}\right).$$

Pour l'estimation sans biais de la variance σ^2, on obtient :

$$\hat{\sigma}^2 = \frac{1}{n-p} \sum_{i=1}^{p} \sum_{k=1}^{n_i} \left(Y_{ik} - \frac{1}{n_i} \sum_{k=1}^{n_i} Y_{ik} \right)^2 = \frac{1}{n-p} \sum_{i=1}^{p} \sum_{k=1}^{n_i} (Y_{ik} - Y_{i.})^2 .$$

On a

$$\frac{(n-p)\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-p).$$

On peut dans ce cadre tester l'effet de la variété en appliquant les résultats de la section 3.2.3. Cela revient à tester H_0 : "$\mu_1 = \mu_2 = \dots = \mu_p$" contre H_1 : "il existe j et j' tels que $\mu_j \neq \mu_{j'}$". On va noter

$$\forall\, i \in \{1, \dots, p\}, \quad Y_{i.} = \frac{1}{n_i} \sum_{k=1}^{n_i} Y_{ik}, \text{ et } Y_{..} = \frac{1}{n} \sum_{i=1}^{p} \sum_{k=1}^{n_i} Y_{ik}.$$

Appliquant le Théorème 3.3, avec

$$X_0 = (1, \dots, 1)^* \in \mathbb{R}^n$$

$$\begin{aligned} F &= \frac{\left(\|X_0(X_0^*X_0)^{-1}X_0^*Y - X(X^*X)^{-1}X^*Y\|^2\right)/(p-1)}{(\|Y - X(X^*X)^{-1}X^*Y\|^2)/(n-p)} \\ &= \frac{\left(\sum_{i=1}^{p} \sum_{k=1}^{n_i} (Y_{i.} - Y_{..})^2\right)/(p-1)}{\left(\sum_{i=1}^{p} \sum_{k=1}^{n_i} (Y_{ik} - Y_{i.})^2\right)/(n-p)}, \end{aligned}$$

et sous H_0,

$$F \sim \mathcal{F}(p-1, n-p).$$

Pour $\alpha \in [0,1]$, on note $f_{p-1,n-p,1-\alpha}$ le quantile d'ordre $1-\alpha$ de la loi de Fisher à $p-1$ et $n-p$ degrés de liberté. Le test

$$\phi(Y) = 1_{F > f_{p-q,n-p,1-\alpha}}$$

est un test de taille α de H_0 contre H_1.

3.4.2 Analyse de la variance à deux facteurs

Pour $I \in \mathbb{N}^*$ et $J \in \mathbb{N}^*$, le **modèle d'analyse de la variance à deux facteurs** s'écrit sous la forme :

$$\forall\, i \in \{1,\dots,I\}, \quad \forall\, j \in \{1,\dots,J\}, \quad \forall\, k \in \{1,\dots,n_{ij}\}$$
$$Y_{ijk} = \mu_{ij} + \varepsilon_{ijk}, \quad \varepsilon_{ijk} \sim \mathcal{N}(0,\sigma^2)$$

où les variables ε_{ijk} sont indépendantes et pour tout $i \in \{1,\dots,I\}$ et pour tout $j \in \{1,\dots,J\}$, $n_{ij} \in \mathbb{N}^*$. Pour simplifier dans la suite, on va supposer que n_{ij} ne dépend pas de i et j et vaut $K \in \mathbb{N}^*$. Par exemple, on explique le rendement associé à la culture du blé en fonction de deux **facteurs** : la variété et l'exposition. On fait pousser sur K parcelles la variété numéro i à l'exposition J et on observe le rendement associé. C'est le vecteur $Y_{ij} = (Y_{ij1},\dots,Y_{ijK})^*$. On écrit sans peine cette modélisation sous la forme (3.2). Dans cette section, pour mieux analyser l'influence des deux facteurs, on va considérer la décomposition de μ_{ij} suivante :

$$\forall\, i \in \{1,\dots,I\}, \quad \forall\, j \in \{1,\dots,J\}, \quad \mu_{ij} = \mu + a_i + b_j + c_{ij},$$

avec pour conserver l'identifiabilité du modèle :

$$\sum_{i=1}^{I} a_i = 0, \quad \sum_{j=1}^{J} b_j = 0,$$

et

$$\forall\, j \in \{1,\dots,J\}, \quad \sum_{i=1}^{I} c_{ij} = 0, \quad \forall\, i \in \{1,\dots,I\}, \quad \sum_{j=1}^{J} c_{ij} = 0.$$

On prend donc $I \geq 2$, $J \geq 2$ et on pose $n = IJK$. Les coefficients a_i représentent l'effet principal du facteur i, les coefficients b_j représentent l'effet principal du facteur j et les coefficients c_{ij} représentent l'interaction entre les facteurs i et j. Dans la suite, on note $\forall\, i \in \{1,\dots,I\}, \quad \forall\, j \in \{1,\dots,J\}$,

$$Y_{...} = \frac{1}{IJK}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} Y_{ijk}, \quad Y_{ij.} = \frac{1}{K}\sum_{k=1}^{K} Y_{ijk}$$

et

$$Y_{i..} = \frac{1}{JK}\sum_{j=1}^{J}\sum_{k=1}^{K} Y_{ijk}, \quad Y_{.j.} = \frac{1}{IK}\sum_{i=1}^{I}\sum_{k=1}^{K} Y_{ijk}.$$

On va d'abord tester si les interactions entre les deux facteurs existent ou non. On va donc tester H_0^1 : "tous les coefficients c_{ij} sont nuls" contre H_1^1 : "il existe un coefficient c_{ij} non nul" avec la variable de test

$$F^1 = \frac{\left(\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(Y_{ij.} - Y_{i..} - Y_{.j.} + Y_{...})^2\right)/((I-1)(J-1))}{\left(\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(Y_{ijk} - Y_{ij.})^2\right)/(n-IJ)},$$

et sous H_0^1,

$$F^1 \sim \mathcal{F}((I-1)(J-1), n-IJ).$$

Pour $\alpha \in [0,1]$, on note $f_{(I-1)(J-1),n-IJ,1-\alpha}$ le quantile d'ordre $1-\alpha$ de la loi de Fisher à $(I-1)(J-1)$ et $n-IJ$ degrés de liberté. Le test

$$\phi(Y) = 1_{F^1 > f_{(I-1)(J-1),n-IJ,1-\alpha}}$$

est un test de taille α de H_0^1 contre H_1^1. Si H_0^1 est acceptée, le modèle se ré-écrit :

$$\forall\, i \in \{1,\dots,I\}, \quad \forall\, j \in \{1,\dots,J\}, \quad \forall\, k \in \{1,\dots,K\}$$

$$Y_{ijk} = \mu + a_i + b_j + \varepsilon_{ijk}, \quad \varepsilon_{ijk} \sim \mathcal{N}(0,\sigma^2).$$

On teste l'effet principal du premier facteur. On va donc tester H_0^2 : "tous les coefficients a_i sont nuls" contre H_1^2 : "il existe un coefficient a_i non nul" avec la variable de test

$$F^2 = \frac{\left(\sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^K (Y_{i..} - Y_{...})^2\right)/(I-1)}{\left(\sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^K (Y_{ijk} - Y_{ij.})^2\right)/(n-IJ)},$$

et sous H_0^2,

$$F^2 \sim \mathcal{F}(I-1, n-IJ).$$

Pour $\alpha \in [0,1]$, on note $f_{I-1,n-IJ,1-\alpha}$ le quantile d'ordre $1-\alpha$ de la loi de Fisher à $I-1$ et $n-IJ$ degrés de liberté. Le test

$$\phi(Y) = 1_{F^2 > f_{I-1,n-IJ,1-\alpha}}$$

est un test de taille α de H_0^2 contre H_1^2. On va aussi tester l'effet principal du second facteur. On va donc tester H_0^3 : "tous les coefficients b_j sont nuls" contre H_1^3 : "il existe un coefficient b_j non nul" avec la variable de test

$$F^3 = \frac{\left(\sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^K (Y_{.j.} - Y_{...})^2\right)/(J-1)}{\left(\sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^K (Y_{ijk} - Y_{ij.})^2\right)/(n-IJ)},$$

et sous H_0^3,

$$F^3 \sim \mathcal{F}(J-1, n-IJ).$$

Pour $\alpha \in [0,1]$, on note $f_{J-1,n-IJ,1-\alpha}$ le quantile d'ordre $1-\alpha$ de la loi de Fisher à $J-1$ et $n-IJ$ degrés de liberté. Le test

$$\phi(Y) = 1_{F^2 > f_{J-1,n-IJ,1-\alpha}}$$

est un test de taille α de H_0^3 contre H_1^3.

3.5 Discussion des hypothèses

Dans cette section, nous allons discuter les hypothèses qui sous-tendent la théorie du modèle linéaire. Ces hypothèses portent en fait sur la modélisation du vecteur des erreurs $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^*$. Elles sont en fait au nombre de 4 :

1. $\mathbb{E}(\varepsilon) = 0$,
2. $\forall\, i \in \{1, \ldots, n\}, \quad \mathrm{var}(\varepsilon_i) = \sigma^2$ (indépendant de i),
3. les composantes du vecteur ε sont indépendantes,
4. la loi de ε est gaussienne.

De ces 4 hypothèses, la dernière est la moins importante dès que l'on dispose de suffisamment de données (en pratique, dès que le nombre de données est de l'ordre de deux ou trois dizaines). De plus, beaucoup des résultats vus précédemment restent valables quand cette hypothèse n'est pas vérifiée.

Pour vérifier a posteriori la validité de ces hypothèses, on s'appuie sur des représentations graphiques. Elles sont réalisées en représentant le graphe du **vecteur des résidus** $\hat{\varepsilon} = Y - X\hat{\beta}$ (ou $\hat{\varepsilon} = Y - \hat{m}$). Si notre travail d'estimation est performant, $\hat{\varepsilon}$ est "proche" de ε et c'est sur $\hat{\varepsilon}$ que l'on va vérifier a posteriori nos hypothèses puisque le vecteur ε n'est pas accessible. De manière systématique, on représente le graphe $((X\hat{\beta})_i, \hat{\varepsilon}_i)_i$. Si les 4 hypothèses sont vérifiées, par le Théorème de Cochran, les deux vecteurs $(X\hat{\beta})_i$ et $(\hat{\varepsilon}_i)_i$ sont gaussiens et indépendants, et le nuage obtenu doit se répartir de manière homogène autour de l'axe des abscisses.

Vérification des hypothèses 1 et 2 : $\mathbb{E}(\varepsilon) = 0$, var$(\varepsilon_i) = \sigma^2$. L'hypothèse $\mathbb{E}(\varepsilon) = 0$ signifie que le modèle posé est correct, que l'on n'a pas oublié un terme pertinent (par exemple un terme quadratique dans l'exemple de la régression linéaire multiple). Une manière de vérifier si un régresseur Z n'a pas été omis est de tracer le graphe $(Z_i, \hat{\varepsilon}_i)_i$. L'hypothèse $\mathrm{var}(\varepsilon_i) = \sigma^2$ est dite hypothèse d'homoscédasticité. Si cette hypothèse n'est pas vérifiée, il faut envisager une transformation du vecteur des observations (cf Azaïs et Bardet (p. 63)). Ces deux hypothèses se vérifient en exploitant le graphe $((X\hat{\beta})_i, \hat{\varepsilon}_i)$ qui doit être homogène autour de l'axe des abscisses. Malheureusement ce graphe ne fournit pas toujours une réponse définitive concernant la violation de ces deux premières hypothèses et il peut indiquer la violation d'autres hypothèses.

Vérification de l'hypothèse 3 : indépendance des composantes de ε. Encore une fois, on peut s'appuyer sur le graphe $((X\hat{\beta})_i, \hat{\varepsilon}_i)_i$. Une autre manière de vérifier cette hypothèse est d'effectuer le **test des runs** (voir le texte sur les tests d'indépendance).

Vérification de l'hypothèse 4 : la loi de ε est gaussienne. Une manière de vérifier cette hypothèse serait de pratiquer un test d'adéquation non-paramétrique classique (test du χ^2 ou test de Kolmogorov-Smirnov). Néanmoins, cette approche serait

quelque peu maladroite car ces tests nécessitent l'hypothèse d'indépendance. On s'appuiera préférentiellement sur le tracé de **la droite de Henry (dite encore graphique du QQ-plot)**. Cette méthode sera illustrée en TP.

Chapitre 4

Fonctions de répartition empiriques

Dans ce chapitre, on s'intéresse à l'estimation de la loi d'une variable aléatoire ainsi qu'aux problèmes de tests associés. Pour traiter ces questions, nous allons chercher à estimer la fonction de répartition de cette variable. Nous sommes donc confrontés à un problème de statistique non-paramétrique et les techniques de démonstration seront donc différentes de celles envisagées dans les chapitres précédents. Pour traiter ce problème, on utilisera la notion de fonction de répartition empirique.

4.1 Généralités

Dans cette section, on considère $X = (X_1, \ldots, X_n)$ un n-échantillon. On note F la fonction de répartition de chacune des variables qui composent cet échantillon :

$$\forall\, t \in \mathbb{R}, \quad F(t) = \mathbb{P}(X_i \leq t).$$

C'est cette fonction F que nous allons chercher à estimer en introduisant la fonction de répartition empirique.

Définition 4.1. *La* **fonction de répartition empirique** *associée à cet échantillon est la fonction*

$$\begin{array}{rcl} \mathbb{R} & \longrightarrow & [0,1] \\ t & \longrightarrow & F_n(t) = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} 1_{X_i \leq t} \end{array}$$

Remarque 4.1. *On a :*

$$\forall\, t \in \mathbb{R}, \quad nF_n(t) \sim Bin(n, F(t)).$$

Pour représenter la fonction F_n, on introduit la statistique d'ordre $(X_{(1)}, \ldots, X_{(n)})$ associée à l'échantillon $(X_1, \ldots, X_n)$ définie par

$$\{X_{(1)}, \ldots, X_{(n)}\} = \{X_1, \ldots, X_n\} \text{ et } X_{(1)} \leq \cdots \leq X_{(n)}.$$

On a :

$$\forall\, t \in \mathbb{R}, \quad F_n(t) = \frac{1}{n}\sum_{i=1}^{n} 1_{X_{(i)} \leq t}.$$

Proposition 4.1. *F_n est une fonction en escalier, croissante, continue à droite et admettant une limite à gauche. Elle est discontinue aux points $(X_{(i)})_{i\in\{1,\dots,n\}}$ et constante sur $[X_{(i)}, X_{(i+1)}[$ pour $i \in \{1, \dots, n-1\}$.*

Preuve : évidente. □

Pour tout $t \in \mathbb{R}$, $F_n(t)$ est un estimateur naturel du paramètre $F(t)$:

Proposition 4.2. *On a pour tout $t \in \mathbb{R}$ que $F_n(t)$ est un estimateur sans biais et fortement consistant de $F(t)$. Par ailleurs,*

$$\forall\, t \in \mathbb{R}, \quad \sqrt{n}(F_n(t) - F(t)) \overset{n\to+\infty}{\longrightarrow} \mathcal{N}(0, F(t)(1-F(t))) \text{ en loi.}$$

Preuve : évidente. □

Dans les sections suivantes, nous allons nous intéresser non pas à la convergence ponctuelle (simple) de F_n vers F mais à la convergence uniforme pour résoudre notre problème de statistique non-paramétrique. Notons que la discontinuité de F_n présente des inconvénients théoriques évidents dans l'optique d'estimer F. Néanmoins, comme elle est constante par morceaux, elle est simple à construire en pratique. Dans les sections suivantes, nous aurons besoin de l'outil de la **fonction inverse généralisée** :

$$\forall\, x \in]0,1[, \quad F^{(-1)}(x) = \inf\{t \in \mathbb{R} : \quad F(t) \geq x\}.$$

Proposition 4.3. *(Ouvrard p. 29) La fonction inverse généralisée se confond avec l'inverse de F quand F est bijective. Elle possède les propriétés suivantes.*

- *La monotonie de F entraîne*

$$\forall\, t \in \mathbb{R}, \quad \forall\, x \in]0,1[, \quad F(t) \geq x \iff t \geq F^{(-1)}(x).$$

- *Si $U \sim U([0,1])$ alors $F^{(-1)}(U)$ est une variable aléatoire dont la fonction de répartition est F.*
- *Si Z est une variable aléatoire de fonction de répartition F continue alors $F(Z) \sim U([0,1])$.*

Preuve : évidente. □

4.2 Théorème de Glivenko-Cantelli

La premier résultat de cette section renforce le théorème de la loi forte des grands nombres.

Théorème 4.1. *Soit $(X_i)_{i\geq 1}$ une suite de variables aléatoires i.i.d. de fonction de répartition F. Avec*

$$\forall\, t \in \mathbb{R}, \quad F_n(t) = \frac{1}{n}\sum_{i=1}^{n} 1_{X_i \leq t},$$

on a :

$$\sup_{t\in\mathbb{R}} [|F_n(t) - F(t)|] \overset{n\to+\infty}{\longrightarrow} 0 \ \textit{p.s.}$$

Preuve : En utilisant la Proposition 4.3, on se ramène au cas où F est la fonction de répartition de la loi uniforme sur $[0,1]$. On fixe $\varepsilon > 0$. On fabrique une grille régulière de $[0,1]$ de pas ε ($1/\varepsilon \in \mathbb{N}$), notée $G_\varepsilon = \{t_k, \quad 0 \leq k \leq \varepsilon^{-1}\}$. Pour $t_k \leq t \leq t_{k+1}$,

$$F_n(t) - t \leq F_n(t_{k+1}) - t \leq F_n(t_{k+1}) - t_{k+1} + \varepsilon,$$

$$-\varepsilon + F_n(t_k) - t_k \leq F_n(t) - t.$$

Finalement, comme $F(t) = t$ pour $t \in [0,1]$,

$$\sup_{t\in[0,1]} [|F_n(t) - t|] \leq \sup_{0\leq k\leq \varepsilon^{-1}} |F_n(t_k) - t_k| + \varepsilon.$$

Par la loi forte des grands nombres,

$$\limsup_{n\longrightarrow+\infty} \sup_{t\in[0,1]} [|F_n(t) - t|] \leq \varepsilon \text{ p.s.}$$

En prenant $\varepsilon = N^{-1}$, on a $\mathbb{P}(A_N) = 1$, où

$$A_N = \left\{ \limsup_{n\longrightarrow+\infty} \sup_{t\in[0,1]} [|F_n(t) - t|] \leq \frac{1}{N} \right\}.$$

en faisant tendre N vers l'infini, les A_N étant des événements décroissants, on a $\mathbb{P}(\cap_N A_N) = 1$. Donc

$$\mathbb{P}\left(\limsup_{n\longrightarrow+\infty} \sup_{t\in[0,1]} [|F_n(t) - t|] = 0 \right) = 1$$

et

$$\sup_{t\in[0,1]} [|F_n(t) - t|] \overset{n\to+\infty}{\longrightarrow} 0 \text{ p.s.}$$

□

4.3 Tests de Kolmogorov

Le Théorème de Glivenko-Cantelli est une généralisation de la loi forte des grands nombres au cas non-paramétrique. La généralisation du TCL est donnée par le Théorème 4.2. La statistique introduite dans ce théorème nous permettra de construire un test d'ajustement à une loi (**test de Kolmogorov**). Dans le même esprit, nous construirons aussi un test d'homogénéité (**test de Kolmogorov - Smirnov**).

Définition 4.2. *Pour toutes fonctions de répartition F et G, on définit les statistiques suivantes :*

$$D(F,G) = \sup_{t\in\mathbb{R}} |G(t) - F(t)|,$$

$$D^-(F,G) = \sup_{t\in\mathbb{R}} (G(t) - F(t)),$$

$$D^+(F,G) = \sup_{t\in\mathbb{R}} (F(t) - G(t)).$$

On a :

Proposition 4.4. *On suppose que F_n est construite à partir d'un échantillon dont la fonction de répartition de chacune des variables de cet échantillon est F. Dès que F est continue, les lois respectives des variables $D(F,F_n)$, $D^+(F,F_n)$ et $D^-(F,F_n)$ sont* **libres** *de F (ne dépendent pas de F).*

Preuve : Pour $D(F,F_n)$ on utilise que :

$$D(F,F_n) = \sup_{t\in\mathbb{R}} |F_n(t) - F(t)| \leq \sup_{x\in[0,1]} \left| \frac{1}{n} \sum_{i=1}^{n} 1_{U_i \leq x} - x \right|, \tag{4.1}$$

où les U_i sont i.i.d. de loi uniforme sur $[0,1]$ avec égalité si F est continue et $D(F,F_n)$ est libre de F. Les cas de $D^+(F,F_n)$ et $D^-(F,F_n)$ se traitent de la même manière. □

Le théorème principal de cette section est le suivant.

Théorème 4.2. *On suppose que F_n est construite à partir d'un échantillon dont la fonction de répartition de chacune des variables de cet échantillon est F. Les variables, $\sqrt{n}D(F,F_n)$, $\sqrt{n}D^+(F,F_n)$ et $\sqrt{n}D^-(F,F_n)$ convergent en loi. Bien évidemment, la loi limite ne dépend pas de F. Pour tout $\lambda > 0$,*

$$\mathbb{P}(\sqrt{n}D(F,F_n) \leq \lambda) \overset{n\to+\infty}{\longrightarrow} 1 - 2\sum_{k=1}^{+\infty} (-1)^{k+1} \exp(-2k^2\lambda^2),$$

la convergence de la série est très rapide. Pour tout $\lambda > 0$,

$$\mathbb{P}(\sqrt{n}D^+(F,F_n) \leq \lambda) = \mathbb{P}(\sqrt{n}D^-(F,F_n) \leq \lambda) \overset{n\to+\infty}{\longrightarrow} 1 - \exp(-2\lambda^2).$$

Preuve : admise. □

Il faut noter que le théorème précédent n'est pas indispensable pour construire une région de confiance pour la fonction F. Contrairement à l'approche paramétrique où on se plaçait souvent dans un cadre asymptotique, ici ce n'est pas nécessaire. En effet, en utilisant la Proposition 4.4 (et sa preuve) pour $0 < \alpha < 1$, on extrait de la table de Kolmogorov $\xi_{n,\alpha}$ le quantile de $D(F^U, F_n^U)$ d'ordre $1-\alpha$, où F^U et F_n^U sont les fonctions de répartition et de répartition empirique associées à la loi uniforme sur $[0,1]$. La bande de confiance de niveau $1-\alpha$ est

$$B_{n,\alpha} = \{F : \quad F_n(t) - \xi_{n,\alpha} \leq F(t) \leq F_n(t) + \xi_{n,\alpha} \quad \forall\, t \in \mathbb{R}\}\,.$$

Soit F_0 une fonction de répartition donnée. Supposons que l'on veuille construire un test de niveau α de l'hypothèse $H_0 : F = F_0$ contre $H_1 : F \neq F_0$. On accepte H_0 si $F_0 \in B_{n,\alpha}$, H_1 sinon (**Test de Kolmogorov**). La puissance du test tend vers 1 quand n tend vers $+\infty$.
De la même manière, on utilisera la bande de confiance associée à $D^+(F^U, F_n^U)$ (respectivement $D^-(F^U, F_n^U)$) pour tester $F = F_0$ contre $F > F_0$ (respectivement pour tester $F = F_0$ contre $F < F_0$). Ici $F > F_0$ signifie, par exemple, qu'il existe $t \in \mathbb{R}$ tel que $F(t) > F_0(t)$.
Dans le même esprit, nous allons construire un test d'homogénéité. On observe deux échantillons de taille respective n et m $X = (X_1, \ldots, X_n)$ et $Y = (Y_1, \ldots, Y_m)$. On veut tester si la loi de chacune des variables X_i est la même que la loi de chacune des variables Y_i. On note F la fonction de répartition de chacune des variables X_i et G la fonction de répartition de chacune des variables Y_i. On veut tester $H_0 : F = G$ contre $F \neq G$. Pour cela, on introduit :

$$\forall\, t \in \mathbb{R}, \quad F_n(t) = \frac{1}{n}\sum_{i=1}^{n} 1_{X_i \leq t}, \quad G_m(t) = \frac{1}{m}\sum_{i=1}^{m} 1_{Y_i \leq t}$$

et on pose

$$D_{m,n} = \sup_{t \in \mathbb{R}} |F_n(t) - G_m(t)|\,.$$

On a le résultat suivant.

Proposition 4.5. *Sous l'hypothèse $H_0 : F = G$, la statistique $D_{m,n}$ est libre de F et G. Pour $\alpha \in [0,1]$ et avec $H_1 : F \neq G$, on construit un test de niveau exactement α de H_0 contre H_1 en considérant*

$$\phi(X, Y) = 1_{D_{m,n} > d_{m,n,\alpha}}$$

où $d_{m,n,\alpha}$ ne dépend que de m, n et α. La puissance du test tend vers 1 quand m ou n tendent vers $+\infty$.

Preuve : évidente. □

Il faut noter que dans ce test, la connaissance de F et G n'est pas nécessaire. De la même manière, on construit un test de $F = G$ contre $F > G$ en posant

$$\phi(X, Y) = 1_{D^+_{m,n} > d^+_{m,n,\alpha}}$$

où

$$D^+_{m,n} = \sup_{t \in \mathbb{R}} (F_n(t) - G_m(t))$$

et $d^+_{m,n,\alpha}$ ne dépend que de m, n et α. Donnons pour finir deux exemples de tests d'appartenance à une famille de lois en s'inspirant de ce qui précède.

Test d'appartenance à la famille des lois gaussiennes. On se donne un n-échantillon $X = (X_1, \ldots, X_n)$. On veut tester H_0 : la loi de chacune des variables X_i est gaussienne contre H_1 : la loi de chacune des variables X_i n'est pas gaussienne. Pour cela, on note Φ la fonction de répartition de la loi $\mathcal{N}(0, 1)$. Si H_0 est vraie et si F est la fonction de répartition de chacune des variables X_i, alors

$$\exists\, (\mu, \sigma) \in \mathbb{R} \times \mathbb{R}^*_+, \quad \forall\, t \in \mathbb{R}, \quad F(t) = \Phi\left(\frac{t - \mu}{\sigma}\right),$$

donc $\Phi^{-1}(F)$ est affine. Au niveau α, on accepte donc H_0 si et seulement si il existe une droite affine appartenant à

$$\Phi^{-1}(B_{n,\alpha}) = \left\{ f : \quad \Phi^{-1}(F_n(t) - \xi_{n,\alpha}) \leq f(t) \leq \Phi^{-1}(F_n(t) + \xi_{n,\alpha}) \quad \forall\, t \in \mathbb{R} \right\}.$$

Test d'appartenance à la famille des lois exponentielles. On se donne un n-échantillon $X = (X_1, \ldots, X_n)$. On veut tester H_0 : la loi de chacune des variables X_i est exponentielle contre H_1 : la loi de chacune des variables X_i n'est pas exponentielle. Si H_0 est vraie et si F est la fonction de répartition de chacune des variables X_i, alors

$$\exists\, \lambda \in \mathbb{R}^*_+, \quad \forall\, t \in \mathbb{R}, \quad F(t) = (1 - \exp(-t/\lambda))\, 1_{t>0}.$$

En notant

$$\forall\, t \in \mathbb{R}, \quad F_n(t) = \frac{1}{n} \sum_{i=1}^{n} 1_{X_i \leq t}, \quad F_{\overline{X}_n}(t) = \left(1 - \exp(-t/\overline{X}_n)\right) 1_{t>0},$$

on introduit

$$E_n = \sup_{t \in \mathbb{R}} \left| F_n(t) - F_{\overline{X}_n}(t) \right|$$

qui est une statistique libre de λ. On construit alors comme précédemment un test de H_0 contre H_1.

Tous les tests introduits dans cette section sont basés sur des statistiques où un sup intervient. Il faut constater que pour chaque cas, ce sup peut être remplacé par un max. Il n'y a donc pas de problèmes d'un point de vue pratique.

4.4 Estimation de quantiles

Dans les chapitres précédents, on a vu qu'il est très important de connaître les quantiles des statistiques qui nous permettent de construire des régions de confiance et des tests. On va voir dans cette section que l'on peut estimer efficacement les quantiles de n'importe quelle loi à l'aide de la fonction de répartition empirique. Rappelons la définition suivante.

Définition 4.3. *Soit $q \in [0,1]$. On appelle* **quantile d'ordre q de la loi $\mathbb{P}$** *la quantité*

$$z_q = \inf \{x : \quad F(x) \geq q\},$$

où F est la fonction de répartition associée à la loi $\mathbb{P}$.

Définissons le quantile empirique d'une loi.

Définition 4.4. *Soit $q \in [0,1]$. On appelle* **quantile empirique d'ordre q de $\mathbb{P}$** *la quantité*

$$\hat{z}_{n,q} = \inf \left\{ x : \quad F_n(x) = \frac{1}{n} \sum_{i=1}^{n} 1_{X_i \leq x} \geq q \right\},$$

où $(X_1, \ldots, X_n)$ est un n-échantillon de loi $\mathbb{P}$.

On a le résultat suivant.

Théorème 4.3. *Soit $0 < q < 1$ est tel que z_q est l'unique solution de $F(x^-) \leq q \leq F(x)$ (pas de plat au voisinage de q), alors $\hat{z}_{n,q}$ est un estimateur fortement consistant de z_q.*

Preuve : Soit $\varepsilon > 0$.

$$\begin{aligned}
\mathbb{P}(\hat{z}_{n,q} > z_q + \varepsilon) &= \mathbb{P}(q > F_n(z_q + \varepsilon)) \\
&= \mathbb{P}(\sum_{i=1}^{n} 1_{X_i > z_q + \varepsilon} > n(1-q)) \\
&= \mathbb{P}(\sum_{i=1}^{n} 1_{X_i > z_q + \varepsilon} - \mathbb{P}(X_i > z_q + \varepsilon) > n(F(z_q + \varepsilon) - q))
\end{aligned}$$

Utilisons le lemme de Hoeffding.

Lemme 4.1. *Soit $(Y_1, \ldots, Y_n)$ une suite de variables indépendantes telles que $\mathbb{E}(Y_i) = 0$ et pour tout i, $a_i \leq Y_i \leq b_i$ p.s., alors*

$$\forall\, \lambda > 0, \quad \mathbb{P}(\sum_{i=1}^{n} Y_i \geq \lambda) \leq \exp\left(-\frac{2\lambda^2}{\sum_{i=1}^{n} (b_i - a_i)^2} \right).$$

Avec $\delta_\varepsilon = \frac{1}{2}\min(F(z_q+\varepsilon)-q; q-F(z_q-\varepsilon)) > 0$, on conclut que

$$\mathbb{P}(|\hat{z}_{n,q} - z_q| > \varepsilon) \leq 2\exp(-2n\delta_\varepsilon^2).$$

Pour $n_0 \in \mathbb{N}$,

$$\begin{aligned}
\mathbb{P}(\sup_{n\geq n_0} |\hat{z}_{n,q} - z_q| > \varepsilon) &\leq \sum_{n\geq n_0} \mathbb{P}(|\hat{z}_{n,q} - z_q| > \varepsilon) \\
&\leq \sum_{n\geq n_0} 2\exp(-2n\delta_\varepsilon^2) \\
&\leq \frac{2\exp(-2n_0\delta_\varepsilon^2)}{1-\exp(-2\delta_\varepsilon^2)}
\end{aligned}$$

Donc pour tout $p \in \mathbb{N}$, il existe $n_0(p)$ tel que

$$\mathbb{P}(\sup_{n\geq n_0(p)} |\hat{z}_{n,q} - z_q| > \varepsilon) \leq 2^{-p}.$$

Par Borel-Cantelli, presque sûrement, il existe n_0 tel que

$$\sup_{n\geq n_0} |\hat{z}_{n,q} - z_q| \leq \varepsilon.$$

□

Chapitre 5

Transformée de Laplace - grandes déviations

5.1 Introduction

Dans ce cours, nous introduisons la notion de transformée de Laplace. Nous verrons que cet outil est très puissant pour calculer les moments d'une variable aléatoire, caractériser la loi ou établir la convergence en loi de variables aléatoires. Enfin, la transformée de Laplace permettra d'introduire la transformée de Legendre qui est à la base des inégalités de grandes déviations, outil fondamental en probabilités et statistique.

5.2 Transformée de Laplace : Définition et propriétés générales

5.2.1 Cas des variables positives ou négatives

Définition 5.1. *Soit $(\Omega, \mathcal{A}, \mathbb{P})$ un espace probabilisé. Soit $X : \Omega \longrightarrow \mathbb{R}$ une variable aléatoire. On appelle transformée de Laplace de X, la fonction L_X à valeurs dans $\mathbb{R}_+$ telle que*

$$L_X(t) = \mathbb{E}(e^{tX}),$$

pour tout $t \in \mathcal{D}_{L_X}$, avec

$$\mathcal{D}_{L_X} = \{t : \quad \mathbb{E}(e^{tX}) < \infty\}.$$

S'il n'y a pas de risque de confusion, on notera L plutôt que L_X. En considérant le produit scalaire usuel, cette définition s'étend à $\mathbb{R}^d$, $d \geq 1$. Dans ce cours, on se restreindra au cas $d = 1$, même si la plupart des résultats restent vrais dans $\mathbb{R}^d$.

Exemple 5.1. *Si $X \sim \mathcal{E}(\lambda)$, $L_X(t) = \lambda/(\lambda - t)$, $\mathcal{D}_{L_X} =]-\infty, \lambda[$.*

Exemple 5.2. *Si $X \sim \mathcal{N}(0,1)$, $L_X(t) = \exp(t^2/2)$, $\mathcal{D}_{L_X} = \mathbb{R}$.*

Proposition 5.1. *La transformée de Laplace est une fonction convexe. Par ailleurs, $\mathcal{D}_{L_X}$ est un intervalle de $\mathbb{R}$ contenant 0.*

Preuve : se déduit de la convexité de la fonction exponentielle. □

On note que pour tout $t \in \mathcal{D}_{L_X}$, $L_X(t) = \phi_X(-it)$ où ϕ_X est la fonction caractéristique de X. D'autre part, si G_X est la fonction génératrice de la variable X positive discrète, alors pour tout $t \in \mathcal{D}_{L_X}$, $L_X(t) = G_X(\exp(t))$. Un intérêt de la transformée de Laplace repose sur la propriété élémentaire suivante :

Proposition 5.2. *Si X et Y sont deux variables indépendantes, pour tout $t \in \mathcal{D}_{L_X} \cap \mathcal{D}_{L_Y}$,*

$$L_{X+Y}(t) = L_X(t) \times L_Y(t).$$

Preuve : immédiate. □

Passons aux choses sérieuses.

Théorème 5.1. *Si $\mathcal{D}_{L_X}$ est d'intérieur non vide, alors L_X est analytique sur $\overset{o}{\mathcal{D}}_{L_X}$:*

$$\forall\, t_0 \in \overset{o}{\mathcal{D}}_{L_X} \;\; \forall\, t \in]t_0 - r; t_0 + r[\subset \overset{o}{\mathcal{D}}_{L_X}, \quad L_X(t) = \sum_{k=0}^{\infty} \frac{(t-t_0)^k}{k!} \mathbb{E}(X^k \exp(t_0 X)).$$

Preuve : Soit $t_0 \in \overset{o}{\mathcal{D}}_{L_X}$ et r tel que $]t_0 - r; t_0 + r[\subset \overset{o}{\mathcal{D}}_{L_X}$. Soit $t \in]t_0 - r; t_0 + r[$. On pose $u = t - t_0$. Donc $t_0 + u$ et $t_0 - u$ appartiennent à $]t_0 - r; t_0 + r[$.

$$\begin{aligned} L_X(t) &= L_X(t_0 + u) \\ &= \mathbb{E}(\exp((t_0 + u)X)) \\ &= \int \exp(ux) \exp(t_0 x) \mathbb{P}_X(dx) \\ &= \int \lim_{n \to +\infty} \sum_{k=0}^{n} \frac{u^k x^k}{k!} \exp(t_0 x) \mathbb{P}_X(dx). \end{aligned}$$

Pour tout n,

$$\begin{aligned} \left| \sum_{k=0}^{n} \frac{u^k x^k}{k!} \exp(t_0 x) \right| &\leq \sum_{k=0}^{n} \frac{|ux|^k}{k!} \exp(t_0 x) \\ &\leq \exp(|ux|) \exp(t_0 x) \\ &\leq (\exp(ux) + \exp(-ux)) \exp(t_0 x), \end{aligned}$$

et on a

$$\mathbb{E}((\exp(uX) + \exp(-uX)) \exp(t_0 X)) < \infty.$$

Donc

$$\begin{aligned} L_X(t) &= \sum_{k=0}^{\infty} \int \frac{u^k x^k}{k!} \exp(t_0 x) \mathbb{P}_X(dx) \\ &= \sum_{k=0}^{\infty} \frac{(t-t_0)^k}{k!} \mathbb{E}(X^k \exp(t_0 X)). \end{aligned}$$

Remarque 5.1. *On note au passage que* $\mathbb{E}(|X|^k \exp(t_0 X)) < \infty$.

□

Corollaire 5.1. *Si* $\mathcal{D}_{L_X}$ *est d'intérieur non vide, alors*

1. L_X *est* C^∞ *en* $t_0 \in \overset{o}{\mathcal{D}}_{L_X}$ *et*

$$L_X^{(k)}(t_0) = \mathbb{E}(X^k \exp(t_0 X)),$$

2. *si* $0 \in \overset{o}{\mathcal{D}}_{L_X}$, X *possède des moments de tous ordres et* $L_X^{(k)}(0)$ *est le moment d'ordre* k *de* X.

Preuve : immédiate. □

On a une réciproque au théorème :

Théorème 5.2. *Si* μ *est une probabilité admettant des moments de tout ordre* μ_k *et si la série entière*

$$l(z) = \sum_{k=0}^{\infty} \frac{\mu_k}{k!} z^k$$

a un rayon de convergence $R > 0$ *alors* μ *est la seule mesure de probabilité admettant pour moments* μ_1, μ_2,...

Preuve : On a $\mu_k = \int x^k \mu(dx)$ et on note $v_k = \int |x|^k \mu(dx)$. Pour $0 < r < R$, et tout $k \geq 1$,

$$\begin{aligned} \frac{v_{2k-1} r^{2k-1}}{(2k-1)!} &\leq \int (1 + x^{2k}) \mu(dx) \frac{r^{2k-1}}{(2k-1)!} \\ &\leq \frac{r^{2k-1}}{(2k-1)!} + \frac{r^{2k-1} \mu_{2k}}{(2k-1)!} \\ &\leq \frac{r^{2k-1}}{(2k-1)!} + \frac{s^{2k} \mu_{2k}}{(2k)!}, \end{aligned}$$

pour $r < s < R$ et k assez grand. Et on obtient :

$$\lim_{k\to\infty} \frac{v_k r^k}{k!} = 0.$$

Maintenant, on utilise que

$$e^{ix} = \sum_{k=0}^{n} \frac{(ix)^k}{k!} + \frac{i^{n+1}}{n!} \int_0^x (x-s)^n e^{is} ds,$$

et

$$\left| e^{ix} - \sum_{k=0}^{n} \frac{(ix)^k}{k!} \right| \leq \frac{|x|^{n+1}}{(n+1)!}.$$

Enfin, en notant ϕ la fonction caractéristique de μ, on a donc pour tout t et tout h,

$$\left| \phi(t+h) - \sum_{k=0}^{n} \frac{h^k}{k!} \int (ix)^k e^{itx} \mu(dx) \right| \leq \frac{|h|^{n+1} v_{n+1}}{(n+1)!}.$$

Si $|h| \leq r$, on a donc

$$\begin{aligned} \phi(t+h) &= \sum_{k=0}^{\infty} \frac{h^k}{k!} \int (ix)^k e^{itx} \mu(dx) \\ &= \sum_{k=0}^{\infty} \frac{h^k}{k!} \phi^{(k)}(t). \end{aligned}$$

En particulier, pour tout $|h| \leq r$,

$$\phi(h) = \sum_{k=0}^{\infty} \frac{h^k}{k!} i^k \mu_k.$$

Les deux dernières égalités achèvent la démonstration du théorème. □

Corollaire 5.2. *Si* $0 \in \overset{o}{\mathcal{D}}_{L_X}$ *alors les conditions précédentes sont vérifiées et* L_X *détermine la loi de* X *car elle détermine les moments.*

Preuve : immédiate. □

5.2.2 Cas des variables positives

Théorème 5.3. *Si X et Y sont deux variables positives et si $L_X = L_Y$ sur $\mathbb{R}^-$ alors $X \sim Y$.*

Preuve : On se donne $Y_\lambda \sim \mathcal{P}(\lambda)$. L'inégalité de Tchebycheff montre que $\lim_{\lambda\to+\infty} Y_\lambda/\lambda = 1$ en probabilité. La fonction de répartition de Y_λ/λ est

$$G_\lambda(t) = \sum_{k=0}^{[t\lambda]} \exp(-\lambda)\frac{\lambda^k}{k!}.$$

Par ce qui précède,

$$\lim_{\lambda\to+\infty} G_\lambda(t) = \begin{cases} 1 & \text{si } t > 1, \\ 0 & \text{si } t < 1. \end{cases}$$

Le théorème de dérivation sous le signe $\int$ montre que pour $s > 0$,

$$L_X^{(k)}(-s) = \int_0^\infty y^k e^{-sy}\mu_X(dy).$$

Par conséquent, pour $s > 0$, $x \geq 0$,

$$\begin{aligned} \sum_{k=0}^{[sx]} \frac{s^k}{k!} L_X^{(k)}(-s) &= \int_0^{+\infty} \sum_{k=0}^{[sx]} e^{-sy}\frac{(sy)^k}{k!}\mu_X(dy) \\ &= \int_0^{+\infty} G_{sy}(x/y)\mu_X(dy). \end{aligned}$$

Aux points x tels que $\mu_X(x) = 0$, on a donc

$$\lim_{s\to+\infty} \sum_{k=0}^{[sx]} \frac{s^k}{k!} L_X^{(k)}(-s) = \mu_X([0,x]).$$

Ainsi, aux points x tels que $\mu_X(x) = 0$, la valeur de la fonction de répartition de X est déterminée par les valeurs de la transformée de Laplace (on a une formule d'inversion). La continuité à droite de la fonction de répartition et la formule précédente permettent de conclure que ce qui précède est vrai en tout point x de $\mathbb{R}^+$. Cela achève la preuve du théorème. □

Théorème 5.4. *On se donne des variables $(X_n)_{n\in\mathbb{N}}$ et X réelles positives. On note $(L_n)_{n\in\mathbb{N}}$ et L les transformées de Laplace associées. Alors $(X_n)_{n\in\mathbb{N}}$ converge en loi vers X si et seulement si $(L_n)_{n\in\mathbb{N}}$ converge simplement vers L sur $\mathbb{R}^-$.*

Preuve : On note μ_n la mesure de probabilité associée à X_n et μ celle associée à X. Si $X_n \xrightarrow{\mathcal{L}} X$, alors pour tout $t \in \mathbb{R}^-$, comme $x \longrightarrow \exp(tx)$ est continue bornée sur $\mathbb{R}^+$,

$$\lim_{n\to+\infty} L_n(t) = \lim_{n\to+\infty} \int_0^\infty \exp(tx)\mu_n(dx) = \int_0^\infty \exp(tx)\mu(dx) = L(t).$$

Réciproquement, on considère une sous-suite quelconque (μ_{n_k}) qui converge vers une mesure ν. On note que ν est de masse inférieure ou égale à 1. Mais pour tout $t \in \mathbb{R}^-$,

$$\int_0^{+\infty} \exp(tx)\mu(dx) = L(t) = \lim_{k\to+\infty} L_{n_k}(t) = \lim_{k\to+\infty} \int_0^{+\infty} \exp(tx)\mu_{n_k}(dx) = \int_0^{+\infty} \exp(tx)\nu(dx).$$

Avec $t = 0$, on conclut que ν est une mesure de probabilité. Par ailleurs, le théorème précédent nous assure que $\nu = \mu$ et qu'il y a donc unicité de la limite. [Bill] pages 336 et 337 permet de conclure (notion de tension d'une mesure $\longleftrightarrow$ compacité). □

5.3 Exemple en fiabilité

Une machine est contrôlée et remise à neuf à des dates aléatoires et on souhaiterait déterminer la loi de sa durée de vie. Pour cela, on note :

- X_i la durée de vie de la machine près le $i$ème contrôle,
- Y_i la durée qui s'écoule entre le $i$ème et le $(i+1)$ème contrôle.

La durée de vie totale de la machine, notée Z, vérifie

$$Z = Y_1 + Y_2 + \cdots + Y_{N-1} + X_N,$$

où $N = \inf\{k \in \mathbb{N}^* : \quad X_k < Y_k\}$.

Proposition 5.3. *Si les variables $(X_1, Y_1, \dots)$ sont indépendantes, les $(Y_i)_{i\in\mathbb{N}^*}$ i.i.d. non identiquement nulles presque sûrement et les $(X_i)_{i\in\mathbb{N}^*}$ i.i.d. de loi $\mathcal{E}(\lambda)$ alors la loi de Z ne dépend pas de celle des $(Y_i)_{i\in\mathbb{N}^*}$ et c'est encore une loi $\mathcal{E}(\lambda)$.*

Preuve : Comme $Z = Y_1 + Y_2 + \cdots + Y_{N-1} + X_N$,

$$\begin{aligned}
L_Z(t) &= \mathbb{E}(e^{tZ}) \\
&= \mathbb{E}(e^{tY_1} \times e^{tY_2} \times \cdots \times e^{tY_{N-1}} \times e^{tX_N}) \\
&= \sum_{k=1}^{\infty} \mathbb{E}(e^{tY_1} \times e^{tY_2} \times \cdots \times e^{tY_{k-1}} \times e^{tX_k} \times 1_{N=k}) \\
&= \sum_{k=1}^{\infty} \mathbb{E}(e^{tY_1} \times e^{tY_2} \times \cdots \times e^{tY_{k-1}} \times e^{tX_k} \times 1_{Y_1\le X_1} \times 1_{Y_2\le X_2} \times \cdots \times 1_{Y_{k-1}\le X_{k-1}} 1_{Y_k > X_k}) \\
&= \sum_{k=1}^{\infty} (\mathbb{E}(e^{tY_1} 1_{Y_1\le X_1}))^{k-1} \mathbb{E}(e^{tX_1} 1_{Y_1 > X_1}) \\
&= \frac{\mathbb{E}(e^{tX_1} 1_{Y_1 > X_1})}{1 - \mathbb{E}(e^{tY_1} 1_{Y_1\le X_1})}
\end{aligned}$$

On calcule :

$$\mathbb{E}(e^{tY_1}1_{Y_1\leq X_1}) = L_Y(t-\lambda),$$
$$\mathbb{E}(e^{tX_1}1_{Y_1>X_1}) = \frac{\lambda}{\lambda-t}(1-L_Y(t-\lambda)).$$

On obtient

$$L_Z(t) = \frac{\lambda}{\lambda-t} = L_{X_1}(t),$$

donc $Z \sim X_1 \sim \mathcal{E}(\lambda)$.

5.4 Transformée de Legendre

5.4.1 Définitions

Définition 5.2. *On définit la fonction Λ comme le logarithme de la transformée de Laplace de X :*

$$\forall\, t \in \mathcal{D}_{L_X}, \quad \Lambda(t) = \log \mathbb{E}(e^{tX}).$$

On montre aisément :

Proposition 5.4. *La fonction Λ est convexe et elle est strictement convexe dès que X n'est pas constante presque sûrement.*

Preuve : Elle découle de l'inégalité de Hölder : pour tout $\lambda \in]0,1[$,

$$\forall\, (t_1,t_2) \in \mathbb{R}^2, \quad \mathbb{E}e^{\lambda t_1 X}e^{(1-\lambda)t_2 X} \leq \mathbb{E}(e^{t_1 X})^{\lambda}\mathbb{E}(e^{t_2 X})^{1-\lambda}.$$

□

Par le théorème de dérivation sous l'intégrale, on montre :

Proposition 5.5. *La fonction Λ est dérivable sur $\overset{o}{\mathcal{D}}_{L_X}$, de dérivée $t \mapsto \mathbb{E}(Xe^{tX})/\mathbb{E}(e^{tX})$. En particulier, si $0 \in \overset{o}{\mathcal{D}}_{L_X}$, $\Lambda'(0) = \mathbb{E}(X)$.*

Preuve : Voir ce qui précède. □

On définit ainsi la transformée de Legendre de X :

Définition 5.3. (transformée de Legendre ou de Cramer)
Si Λ est le logarithme de la transformée de Laplace de X, la transformée de Legendre de X est la fonction :

$$\forall\, x \in \mathbb{R}, \quad \Lambda^*(x) = \sup_{t\in\mathbb{R}}(tx - \Lambda(t))$$

Remarque 5.2. *Puisque $\Lambda(0) = 0$, Λ^* est à valeurs dans $[0,+\infty]$. De plus,*

$$\Lambda^*(x) = \sup_{t\in\mathcal{D}_{L_X}}(tx - \Lambda(t)).$$

5.4.2 Exemples de transformées de Legendre

- si $X \sim \mathcal{P}(a)$, alors $\Lambda^*(x) = \begin{cases} a - x + x\log(x/a) & \text{si } x > 0 \\ +\infty & \text{sinon} \end{cases}$
- si $X \sim Ber(p)$, alors $\Lambda^*(x) = \begin{cases} x\log(\frac{x}{p}) + (1-x)\log(\frac{1-x}{1-p}) & \text{si } x \in [0,\ 1] \\ +\infty & \text{sinon} \end{cases}$
- si $X \sim \mathcal{E}(a)$, alors $\Lambda^*(x) = \begin{cases} ax - 1 - \log(ax) & \text{si } x > 0 \\ +\infty & \text{sinon} \end{cases}$
- si $X \sim \text{Gamma}(a)$, alors $\Lambda^*(x) = \begin{cases} x - a - a\log(x/a) & \text{si } x > 0 \\ +\infty & \text{sinon} \end{cases}$
- si $X \sim \mathcal{N}(0,1)$ alors $\Lambda^*(x) = x^2/2$
- si X suit une loi de Cauchy, alors $\Lambda^*(x) = 0$

Proposition 5.6. *On a :*
- Λ^* *est convexe.*
- *Si* $X \in \mathbb{L}^1$ *et si* $m = \mathbb{E}(X)$, *alors* $\Lambda^*(m) = 0$. *De plus,* Λ^* *est croissante sur* $[m, +\infty[$, *décroissante sur* $]-\infty, m]$.
- *Si* $0 \in \overset{o}{\mathcal{D}}_{L_X}$,

$$\forall x \geq m, \quad \Lambda^*(x) = \sup_{t \geq 0}(tx - \Lambda(t)), \quad P(X \geq x) \leq e^{-\Lambda^*(x)},$$

$$\forall x \leq m, \quad \Lambda^*(x) = \sup_{t \leq 0}(tx - \Lambda(t)), \quad P(X \leq x) \leq e^{-\Lambda^*(x)}.$$

Preuve :
- Λ^* est convexe comme sup de fonctions convexes.
- $\Lambda(t) = \log \mathbb{E}(e^{tX}) \geq \mathbb{E}(tX) = tm$. Donc $\Lambda^*(m) \leq 0$, donc $\Lambda^*(m) = 0$. Par convexité et positivité, Λ^* est croissante sur $[m, +\infty[$, décroissante sur $]-\infty, m]$.
- On ne traite que le premier cas. On a $\Lambda'(0) = m$. Si $t < 0$, par stricte croissance de Λ', on a : $x - \Lambda'(t) > x - \Lambda'(0) \geq m - \Lambda'(0) = 0$, si $x \geq m$. Donc la fonction $t \longrightarrow tx - \Lambda(t)$ est strictement croissante sur $\mathbb{R}^-$. □

5.4.3 Autres propriétés de la transformée de Legendre

On suppose que $0 \in \overset{\circ}{D}$ et X est une variable aléatoire non constante presque sûrement. On notera $\overset{\circ}{D} =]\theta^-, \theta^+[$. On a :

1. Comme $0 \in \overset{\circ}{D}$, $\Lambda^*(x) \to +\infty$ lorsque $|x| \to +\infty$.
2. Si $D = \mathbb{R}$, alors $\frac{\Lambda^*(x)}{|x|} \to +\infty$ lorsque $|x| \to +\infty$.
3. Soit $x^+ = \lim_{\theta \nearrow \theta^+} \Lambda'(\theta)$ et $x^- = \lim_{\theta \searrow \theta^-} \Lambda'(\theta)$. On a :

$$\forall\ x \in]x^-, x^+[, \quad \Lambda^*(x) = x\,\theta^*(x) - \Lambda(\theta^*(x)),$$

où $\theta^*(x) = (\Lambda')^{-1}(x)$.

5.4.4 Application : calcul d'intervalles de confiance

Théorème 5.5. *Soit $X_1, \ldots, X_n$ un échantillon de même loi que X. On suppose que $0 \in \overset{o}{\mathcal{D}}_{L_X}$. On note m l'espérance de X. On a*

$$\forall\, \varepsilon_1, \varepsilon_2 > 0, \quad \mathbb{P}(\frac{X_1 + \ldots + X_n}{n} - m \notin]-\varepsilon_1, \varepsilon_2[) \leq \exp(-n\Lambda^*(m-\varepsilon_1)) + \exp(-n\Lambda^*(m+\varepsilon_2)),$$

où Λ^* est la transformée de Legendre de X.

Preuve : Il suffit de remarquer :

$$\begin{aligned} \mathbb{P}(\frac{X_1 + \ldots + X_n}{n} - m \leq -\varepsilon_1) &= \mathbb{P}(X_1 + \ldots + X_n \leq n(m - \varepsilon_1)) \\ &\leq \exp(-\Lambda^*(n(m - \varepsilon_1))) \\ &= \exp(-n\Lambda^*(m - \varepsilon_1)). \end{aligned}$$

□

Application : On montre par le calcul que si $X \sim Ber(p)$, alors $\Lambda^*(p + \varepsilon) \geq 2\varepsilon^2$ et $\Lambda^*(p - \varepsilon) \geq 2\varepsilon^2$ (c.f. Inégalité de Hoeffding). On obtient l'intervalle de confiance de niveau α pour p ($\varepsilon_1 = \varepsilon_2 = \sqrt{\frac{1}{2n}\log\left(\frac{2}{\alpha}\right)}$) :

$$\left[\overline{X}_n - \sqrt{\frac{1}{2n}\log\left(\frac{2}{\alpha}\right)}; \overline{X}_n + \sqrt{\frac{1}{2n}\log\left(\frac{2}{\alpha}\right)}\right].$$

A noter, que l'on obtient les mêmes minorations de la transformée de Legendre en utilisant $p(1-p) \leq 0.25$ et un développement limité de Λ^* quand $\varepsilon \to 0$.

5.4.5 Grandes déviations

Si $X_1, \ldots, X_n$ est un échantillon de v.a.r. i.i.d. centrées, les lois des grands nombres nous enseignent que la probabilité de l'événement $|X_1 + \ldots + X_n| \geq na$ tend vers 0 si $a > 0$. C'est un "événement rare". Nous allons donner un théorème qui évalue la probabilité de tels événements pour des échantillons non spécifiquement gaussiens. C'est l'objet du théorème des grandes déviations suivant :

Théorème 5.6. (Grandes déviations)
Soit $X_1, \ldots, X_n$ un échantillon de variables aléatoires réelles suivant la même loi qu'une variable aléatoire X centrée non constante presque sûrement et telle que

$$\mathbb{E}(\exp(\theta|X|)) < \infty, \quad \forall\, \theta \geq 0.$$

Soit $a > 0$. Alors

$$\lim_{n\to+\infty} \frac{1}{n} \log \mathbb{P}\left(\frac{X_1 + \ldots + X_n}{n} \geq a\right) = -\Lambda^*(a),$$

où Λ^* est la transformée de Legendre de X. De manière similaire, si $a < 0$,

$$\lim_{n\to+\infty} \frac{1}{n} \log \mathbb{P}\left(\frac{X_1 + \ldots + X_n}{n} \leq a\right) = -\Lambda^*(a).$$

Preuve : Soit $a > 0$ et $t \geq 0$.

$$\begin{aligned}
\mathbb{P}(\sum_{i=1}^{n} X_i \geq na) &= \int 1_{\sum_{i=1}^n x_i \geq na} d\mathbb{P}_X^n(x_1, \ldots, x_n) \\
&\leq \int \exp(t(\sum_{i=1}^{n} x_i - na)) 1_{\sum_{i=1}^n x_i \geq na} d\mathbb{P}_X^n(x_1, \ldots, x_n) \\
&\leq \exp(n\Lambda(t) - nat) \\
&\leq \exp(-n\Lambda^*(a)).
\end{aligned}$$

La preuve de l'inégalité contraire est très longue et utilise des outils sophistiqués. □

Ce résultat est très puissant quand on cherche à construire des régions de confiance pour l'estimation de la moyenne d'un échantillon. Il donne l'intervalle de confiance asymptotique le plus précis... dès que l'on sait calculer Λ^*.

Remarque 5.3. *Le théorème précédent montre que le comportement asymptotique est caractéristique de la loi des X_i. Il faut noter que si a dépend de n et $a = a(n)$ tend vers 0, le comportement ne retient que les deux premiers moments de la loi des X_i comme pour le théorème central limite (théorème des moyennes déviations, voir Genon-Catalot et Picard).*

Chapitre 6

Mesure de la performance d'un estimateur

6.1 Introduction

On se donne $(\Omega, \mathcal{A}, \mathbb{P}_\theta, \theta \in \Theta)$ un modèle statistique où Θ est un ouvert de $\mathbb{R}$ et $g(\theta)$ une quantité à estimer à partir d'une observation X de loi $\mathbb{P}_\theta$. Souvent, on dispose d'un n-échantillon et dans ce cas, $\mathbb{P}_\theta$ dépend de n. Il a été vu précédemment que les bonne propriétés d'un estimateur sont les suivantes :

- être sans biais (ou asymptotiquement sans biais),
- la consistance (forte ou faible),
- la normalité asymptotique.

Il n'est pas rare de trouver plusieurs estimateurs qui possèdent une voire toutes ces propriétés. Lequel alors choisir ? Fisher affirmait que la classe des estimateurs "intéressants" sont ceux qui possèdent au moins toutes ces propriétés. Il conjecturait également que le maximum de vraisemblance fait partie de cette classe et que c'est pour cet estimateur que la variance de la loi limite est minimale. Dans la suite, avant de considérer la conjecture de Fisher, on va introduire la notion de **risque** pour mesurer de manière naturelle la performance d'un estimateur. En utilisant cette notion, on pourra alors étudier l'optimalité de méthodes d'estimation.

6.2 Risque d'un estimateur

Définition 6.1. *Pour chaque estimateur $T(X)$ de $g(\theta) \in \mathbb{R}$ le risque associé est la fonction*

$$\begin{aligned} R(T,.): \ \Theta &\longrightarrow \overline{\mathbb{R}}_+ \\ \theta &\longrightarrow \mathbb{E}_\theta (T(X) - g(\theta))^2 \end{aligned}$$

La notion de risque peut être étendue à des pertes différentes de la perte L_2. Bien entendu un estimateur est d'autant meilleur que son risque est le plus faible sur Θ. En

général, il n'existe pas d'estimateurs uniformément meilleurs sur Θ.

Exemple 6.1. *On considère :*
- $(X_1, \dots, X_n)$ *est un échantillon de loi de Bernoulli de paramètre* $\theta \in \Theta =]0,1[$,
- $g(\theta) = \theta$.

1. Si $T_1 = \overline{X}_n = \frac{S_n}{n}$, $R(T_1, \theta) = \frac{\theta(1-\theta)}{n}$.
2. Si $T_2 = \frac{S_n + a}{n+b}$, $R(T_2, \theta) = \frac{n}{(n+b)^2}\theta(1-\theta) + \frac{(a-b\theta)^2}{(n+b)^2}$.
3. Si $T_3 = \theta_0$, $R(T_3, \theta) = (\theta - \theta_0)^2$.

Conclusion : il faut se donner d'autres critères que le risque pur. Trois solutions parmi d'autres :

1. Ne considérer que les estimateurs **admissibles**. Un estimateur T sera dit **inadmissible** s'il existe T' tel que
$$\forall\, \theta \in \Theta,\ R(T', \theta) \leq R(T, \theta),$$
$$\exists\, \theta_0 \in \Theta,\ R(T', \theta_0) < R(T, \theta_0).$$
2. Choisir des estimateurs qui minimisent une fonction du risque.
Exemple : $\sup_{\theta \in \Theta} R(T, \theta)$. On choisit un estimateur qui minimise cette quantité (**estimateur minimax**).
3. Restreindre la classe des estimateurs. Classiquement, on se restreint à la classe des estimateurs **sans biais**. Rappelons que le biais d'un estimateur intégrable $T(X)$ est $b(\theta) = \mathbb{E}_\theta T(X) - g(\theta)$. Il est dit sans biais si $b(\theta) = 0\ \forall\, \theta \in \Theta$.

On va privilégier cette dernière solution dans la suite. Notons que cela permet d'éliminer la classe des estimateurs constants dès que g est non constant sur Θ. Néanmoins, se restreindre à la classe des estimateurs sans biais présente certains inconvénients. Tout d'abord, si T est sans biais pour l'estimation de $g(\theta)$, cette propriété n'est plus forcément vérifiée pour l'estimation de $h(g(\theta))$ par $h(T)$. Plus ennuyeux, cette classe peut être très restreinte :

Exemple 6.2. *On observe :*
- S_n *suit une loi binomiale de paramètres* $n \in \mathbb{N}^*$ *et* $\theta \in \Theta =]0,1[$,
- $g(\theta) = \theta$.

L'estimateur $T(S_n) = \frac{S_n}{n}$ convient et c'est le seul. En fait, si $T'(S_n)$ est aussi un estimateur sans biais, alors, en posant $h = T - T'$, on a :

$$\begin{array}{lll}
 & \sum_{k=0}^{n} h(k) C_n^k \theta^k (1-\theta)^{n-k} = 0 & \forall\, \theta \in \Theta \\
\Leftrightarrow & \sum_{k=0}^{n} h(k) C_n^k \theta^k \sum_{l=0}^{n-k} C_{n-k}^l (-1)^l \theta^l = 0 & \forall\, \theta \in \Theta \\
\Leftrightarrow & \sum_{m=0}^{n} \theta^m \sum_{k,l:\ l+k=m} C_n^k C_{n-k}^l (-1)^l h(k) = 0 & \forall\, \theta \in \Theta \\
\Leftrightarrow & \sum_{k,l:\ l+k=m} C_n^k C_{n-k}^l (-1)^l h(k) = 0, \quad \forall\, m \in \{0, \dots, n\} & \forall\, \theta \in \Theta \\
\Leftrightarrow & h = 0.
\end{array}$$

La classe des estimateurs sans biais peut même être vide :

Exemple 6.3. *On observe :*
- *S_n suit une loi binomiale de paramètres $n \in \mathbb{N}^*$ et $\theta \in \Theta =]0,1[$,*
- *$g(\theta) = \sqrt{\theta}$.*

La définition que l'on a adoptée du risque permet d'obtenir la **décomposition biais-variance** : Pour tout estimateur $T(X)$ et tout $\theta \in \Theta$,

$$\begin{aligned} R(T,\theta) &= \mathrm{var}_\theta(T(X)) + (b(\theta))^2 \\ &= \mathrm{var}_\theta(T(X)) \end{aligned}$$

si $b(\theta) = 0$.
Conclusion : Dans la suite, on va donc s'intéresser à la recherche d'estimateurs qui minimisent uniformément la variance parmi les estimateurs sans biais de $g(\theta)$. De tels estimateurs seront appelés **UMVU** (Uniform Minimum Variance Unbiased). On a le résultat important suivant :

Théorème 6.1. *Si un estimateur UMVU de $g(\theta)$ existe alors il est unique $\mathbb{P}_\theta$-p.s. pour tout $\theta \in \Theta$.*

Preuve : Soient T_1 et T_2 deux estimateurs UMVU. On pose $U = T_1 - T_2$. Soit $\lambda \in \mathbb{R}$ et $T_1 + \lambda U$, un estimateur sans biais de $g(\theta)$. On a $\forall\, \theta \in \Theta$,

$$\mathbb{E}_\theta(T_1 - g(\theta))^2 \leq \mathbb{E}_\theta(T_1 + \lambda U - g(\theta))^2$$

et donc

$$2\lambda \mathbb{E}_\theta(T_1 U) + \lambda^2 \mathbb{E}_\theta(U^2) \geq 0.$$

Ceci est vrai pour tout $\lambda \in \mathbb{R}$, donc

$$\forall\, \theta \in \Theta, \quad \mathbb{E}_\theta(T_1 U) = 0.$$

De même

$$\forall\, \theta \in \Theta, \quad \mathbb{E}_\theta(T_2 U) = 0$$

et

$$\forall\, \theta \in \Theta, \quad \mathbb{E}_\theta(T_1 - T_2)^2 = \mathbb{E}_\theta(T_1 U) - \mathbb{E}_\theta(T_2 U) = 0.$$

Donc

$$\forall\, \theta \in \Theta, \quad T_1 = T_2 \ \mathbb{P}_\theta\text{-p.s.}$$

6.3 Estimation optimale

Dans toute cette section, on supposera donnée l'observation X de loi $\mathbb{P}_\theta$. On s'intéresse toujours à l'estimation de $g(\theta)$, avec g dérivable sur Θ. On suppose que $\mathbb{P}_\theta$ admet une densité par rapport à la mesure de Lebesgue sur $\mathbb{R}$ (ou la mesure de comptage sur $\mathbb{Z}$) que l'on note $f(x,\theta)$. On suppose que $\theta \longrightarrow \frac{\partial}{\partial\theta} f(x,\theta)$ existe et est continue sur Θ et on fait l'hypothèse suivante :

Si U est une fonction telle que $U(X)$ admet un moment d'ordre 2 par rapport à $\mathbb{P}_\theta$ pour tout $\theta \in \Theta$, alors

$$\forall\, \theta \in \Theta, \quad \frac{\partial}{\partial \theta} \int U(x) f(x,\theta) dx = \int U(x) \frac{\partial}{\partial \theta} f(x,\theta) dx \tag{6.1}$$

6.3.1 Information de Fisher

Définition 6.2. *On définit l'information de Fisher comme*

$$\forall\, \theta \in \Theta,\ I(\theta) = \mathbb{E}_\theta \left(\frac{\partial}{\partial \theta} \log f(X,\theta) \right)^2 .$$

Remarque 6.1. *$I(\theta)$ appartient à l'intervalle $[0, +\infty]$.*

Proposition 6.1. *Supposons que*

$$\forall\, \theta \in \Theta, \quad \mathbb{E}_\theta \left| \frac{\partial}{\partial \theta} \log f(X,\theta) \right| < \infty.$$

Alors

$$\forall\, \theta \in \Theta, \quad I(\theta) = var_\theta \left(\frac{\partial}{\partial \theta} \log f(X,\theta) \right).$$

Preuve : Il suffit d'observer que

$$\begin{aligned}
\mathbb{E}_\theta \frac{\partial}{\partial \theta} \log f(x,\theta) &= \int \left(\frac{\partial}{\partial \theta} \log f(x,\theta) \right) f(x,\theta) dx \\
&= \int \left(\frac{\partial}{\partial \theta} f(x,\theta) \right) \frac{1}{f(x,\theta)} f(x,\theta) dx \\
&= \int \left(\frac{\partial}{\partial \theta} f(x,\theta) \right) dx \\
&= \frac{\partial}{\partial \theta} \int f(x,\theta) dx \\
&= 0.
\end{aligned}$$

□

Exemple 6.4. *Supposons que l'on observe $X \sim \mathcal{P}(\theta)$, $\theta \in \Theta$. Alors, $f(x,\theta) = \exp(-\theta)\frac{\theta^x}{x!}$, $x \in \mathbb{N}$, et $\frac{\partial}{\partial \theta} \log f(x,\theta) = \frac{x}{\theta} - 1$. On obtient alors*

$$I(\theta) = \frac{1}{\theta}.$$

Exemple 6.5. *Supposons que l'on observe $X = (X_1, \ldots, X_n)$ un n-échantillon de loi $\mathcal{N}(\theta, \sigma^2)$, σ connu.*

$$I(\theta) = \frac{n}{\sigma^2}.$$

6.3.2 Borne de Cramer Rao

Le théorème suivant est le résultat essentiel de cette section.

Théorème 6.2. *Soit T une fonction de telle sorte que $T(X)$ est un estimateur sans biais de $g(\theta)$. Alors, si*

$$\forall\, \theta \in \Theta, \quad \mathbb{E}_\theta \left| \frac{\partial}{\partial \theta} \log f(X, \theta) \right| < \infty,$$

$$\forall\, \theta \in \Theta, \quad var_\theta(T(X)) \geq \frac{(g'(\theta))^2}{I(\theta)}.$$

Preuve : Soit $\theta \in \Theta$. Si $\mathbb{E}_\theta(T(X)^2) = +\infty$, il n'y a rien à montrer. Supposons donc que $\mathbb{E}_\theta(T(X)^2) < +\infty$. Utilisant l'hypothèse (6.1), on a :

$$\begin{aligned} g'(\theta) &= \int T(x) \frac{\partial}{\partial \theta} f(x, \theta) dx \\ &= \int T(x) \left(\frac{\partial}{\partial \theta} \log f(x, \theta) \right) f(x, \theta) dx. \end{aligned}$$

En utilisant la Proposition 6.1,

$$g'(\theta) = \mathbb{E}_\theta \left(\frac{\partial}{\partial \theta} \log f(X, \theta) - \mathbb{E}_\theta \left(\frac{\partial}{\partial \theta} \log f(X, \theta) \right) \right) (T(X) - \mathbb{E}_\theta(T(X))),$$

et par l'inégalité de Cauchy-Schwarz,

$$|g'(\theta)| \leq \sqrt{\mathrm{var}_\theta \left(\frac{\partial}{\partial \theta} \log f(X, \theta) \right)} \times \sqrt{\mathrm{var}_\theta(T(X))}.$$

On conclut en appliquant la proposition précédente. □

Le membre de droite de l'inégalité obtenue dans le théorème précédent est appelé **borne de Cramer Rao**. Cette borne donne une minoration de la variance de $T(X)$ et donc du risque de $T(X)$. Un estimateur $T(X)$ qui vérifie

$$\forall\, \theta \in \Theta, \quad \mathrm{var}_\theta(T(X)) = \frac{(g'(\theta))^2}{I(\theta)}$$

est donc un estimateur UMVU. Prolongeons le résultat du Théorème 6.2 :

Proposition 6.2. *Supposons que l'on observe l'échantillon $X = (X_1, \ldots, X_n)$. Alors, si $I_1(\theta)$ est l'information de Fisher de X_1, on a sous les hypothèses du Théorème 6.2,*

$$\forall\, \theta \in \Theta, \quad I(\theta) = nI_1(\theta)$$

et

$$\forall\, \theta \in \Theta, \quad var_\theta(T(X)) \geq \frac{(g'(\theta))^2}{nI_1(\theta)}.$$

Preuve : Evidente. □

Reprenons les Exemples 6.4 et 6.5. L'observation d'un n-échantillon $X = (X_1, \ldots, X_n)$ permet la construction de l'estimateur du maximum de vraisemblance $\overline{X}_n$ (que ce soit pour le modèle de la loi de Poisson ou le modèle de la loi normale) qui est donc un estimateur UMVU. Est-ce un hasard ? La section suivante va nous montrer que non.

6.3.3 Modèles exponentiels

Introduisons les modèles exponentiels à travers le résultat suivant.

Théorème 6.3. *On suppose que l'hypothèse (6.1) est vérifiée et que*

$$\forall\, \theta \in \Theta, \quad O_\theta = \{x : f(x,\theta) > 0\}$$

ne dépende pas de θ. Alors si $T(X)$ est un estimateur non trivial sans biais de $g(\theta)$ et que la minoration du Théorème 6.2 est atteinte $\forall\, \theta \in \Theta$, le modèle statistique est un **modèle exponentiel associé à** *T, c'est-à-dire qu'il existe des fonctions h, a et b telles que*

$$\forall\, \theta \in \Theta, \forall\, x, \quad f(x,\theta) = h(x)\exp(a(\theta)T(x) - b(\theta)).$$

Preuve : Si la minoration du Théorème 6.2 est atteinte $\forall\, \theta \in \Theta$, alors il existe deux fonctions a^* et b^* telles que $\forall\, \theta \in \Theta$,

$$A_\theta = \left\{ x : \quad \frac{\partial}{\partial\theta} \log f(x,\theta) = a^*(\theta)T(x) - b^*(\theta) \right\}$$

vérifie $\mathbb{P}_\theta(A_\theta) = 1$. On a donc

$$\int 1_{A_\theta^c \cap O_\theta} f(x,\theta)dx = 0.$$

Comme $f(x,\theta) > 0\ \forall\, x \in O_\theta$, si μ est la mesure de Lebesgue sur $\mathbb{R}$ ou de comptage sur $\mathbb{Z}$,

$$\mu(A_\theta^c \cap O_\theta) = 0.$$

Comme O_θ ne dépend pas de θ,

$$\forall\, \theta' \in \Theta, \quad \mathbb{P}_{\theta'}(A_\theta^c) = \int 1_{A_\theta^c \cap O_\theta} f(x,\theta')dx = 0.$$

Donc $\mathbb{P}_{\theta'}(A_\theta) = 1\ \forall\, \theta \in \Theta,\ \forall\, \theta' \in \Theta$. En considérant $(\theta_n)_{n\in\mathbb{N}}$ une suite dense de Θ, avec

$$A^* = \cap_{n\in\mathbb{N}} A_{\theta_n},$$

on a $\mathbb{P}_{\theta'}(A^*) = 1\ \forall\ \theta' \in \Theta$. Soient x_1 et x_2 dans A^* tels que $T(x_1) \neq T(x_2)$ (possible), on montre que a^* et b^* sont combinaisons linéaires de $\frac{\partial}{\partial\theta}\log f(x_1,.)$ et $\frac{\partial}{\partial\theta}\log f(x_2,.)$ donc sont continues sur Θ et

$$A^{**} = \cap_{\theta\in\Theta} A_\theta$$

vérifie $\mathbb{P}_{\theta'}(A^{**}) = 1\ \forall\, \theta' \in \Theta$. On conclut en intégrant. □

Donnons des exemples de modèles exponentiels.

Exemple 6.6. *La loi de Poisson de paramètre $\theta > 0$:*

$$\forall\, x \in \mathbb{N}, \quad f(x,\theta) = \exp(-\theta)\frac{\theta^x}{x!} = \frac{1}{x!}\exp(x\log\theta - \theta).$$

Exemple 6.7. *La loi binomiale de paramètres $n \in \mathbb{N}^*$ et $\theta \in]0,1[$:*

$$\forall\, x \in \{0,\dots,n\}, \quad f(x,\theta) = C_n^x \theta^x (1-\theta)^{n-x} = C_n^x \exp\left(x\log\left(\frac{\theta}{1-\theta}\right) + n\log(1-\theta)\right).$$

Exemple 6.8. *La loi normale de moyenne $\theta \in \mathbb{R}$ et de variance σ^2 connue.*

$$\forall\, x \in \mathbb{R}, \quad f(x,\theta) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\theta)^2}{2\sigma^2}\right) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{x^2}{2\sigma^2}\right)\exp\left(\frac{x\theta}{\sigma^2} - \frac{\theta^2}{2\sigma^2}\right).$$

Exemple 6.9. *La loi normale de moyenne μ connue et de variance $\theta > 0$.*

$$\forall\, x \in \mathbb{R}, \quad f(x,\theta) = \frac{1}{\sqrt{2\pi}\sqrt{\theta}}\exp\left(-\frac{(x-\mu)^2}{2\theta}\right) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{\theta}\frac{(x-\mu)^2}{2} - \frac{1}{2}\log(\theta)\right).$$

Les modèles exponentiels fournissent naturellement des estimateurs UMVU :

Théorème 6.4. *Supposons que l'observation X ait une loi $\mathbb{P}_\theta$ associé au modèle exponentiel suivant :*

$$\forall\, \theta \in \Theta, \forall\, x, \quad f(x,\theta) = h(x)\exp(a(\theta)T(x) - b(\theta)),$$

avec
- *$T(X)$ admet un moment d'ordre 2 par rapport à $\mathbb{P}_\theta$, $\forall\, \theta \in \Theta$,*
- *a et b 2 fois dérivables sur Θ tel que a' est non nul sur Θ.*

Alors l'estimateur $T(X)$ est un estimateur UMVU de $g(\theta) = \frac{b'(\theta)}{a'(\theta)}$.

Preuve : On prouve tout d'abord que l'hypothèse (6.1) est vérifiée. Soit U une fonction telle que $U(X)$ admet un moment d'ordre 2 par rapport à $\mathbb{P}_\theta$ pour tout $\theta \in \Theta$. Soit $\theta \in \Theta$ et V un voisinage de θ inclus dans K un compact de Θ.

$$U(x)\frac{\partial}{\partial\theta}f(x,\theta) \;=\; U(x)h(x)(a'(\theta)T(x) - b'(\theta))\exp(a(\theta)T(x) - b(\theta))$$

Mais il existe θ_* et θ^* dans K tels que $\forall\, \theta \in V$, $a(\theta) \in [a(\theta_*), a(\theta^*)]$ et $\forall\, \theta \in V$, $\forall\, x$

$$\begin{aligned} h(x)\exp(a(\theta)T(x) - b(\theta)) &\leq h(x)\exp(a(\theta_*)T(x) - b(\theta)) + h(x)\exp(a(\theta^*)T(x) - b(\theta)) \\ &\leq f(\theta_*,x)\exp(b(\theta_*) - b(\theta)) + f(\theta^*,x)\exp(b(\theta^*) - b(\theta)) \\ &\leq M(f(\theta_*,x) + f(\theta^*,x)), \end{aligned}$$

où M est une constante. Donc,

$$\left|U(x)\frac{\partial}{\partial\theta}f(x,\theta)\right| \leq M'\left(|U(x)T(x)| + |U(x)|\right)(f(\theta_*,x) + f(\theta^*,x)),$$

où M' est une constante. Donc, comme $U(X)$ et $T(X)$ admettent un moment d'odre 2 par rapport $\mathbb{P}_{\theta_*}$ et $\mathbb{P}_{\theta^*}$, le membre de droite est intégrable, le théorème de dérivation sous le signe $\int$ s'applique et on a :

$$\frac{\partial}{\partial\theta}\int U(x)f(x,\theta)dx = \int U(x)\frac{\partial}{\partial\theta}f(x,\theta)dx.$$

L'hypothèse (6.1) est vérifiée. On prouve que $T(X)$ est sans biais en utilisant :

$$0 = \int \frac{\partial}{\partial\theta}f(x,\theta)dx = a'(\theta)\mathbb{E}_\theta(T(X)) - b'(\theta).$$

On conclut la preuve en utilisant la Proposition 6.1 qui établit que

$$I(\theta) = a'(\theta)^2 \mathrm{var}_\theta(T(X))$$

et l'égalité suivante :

$$\begin{aligned} g'(\theta) &= \int (a'(\theta)T^2(x) - b'(\theta)T(x))f(x,\theta)dx \\ &= a'(\theta)\mathbb{E}_\theta(T^2(X)) - b'(\theta)\mathbb{E}_\theta(T(X)) \\ &= a'(\theta)\mathrm{var}_\theta(T(X)). \end{aligned}$$

□

Le rultat précédent s'applique sans difficulté aux Exemples 6.6, 6.7, 6.8 et 6.9. Dans des modèles plus compliqués que ceux étudiés précédemment, la borne de Cramer Rao est rarement atteinte par les estimateurs usuels pour l'estimation de $g(\theta) = \theta$. Néanmoins, considérer le cadre asymptotique associé à la convergence en loi permet de surmonter ce problème :

Théorème 6.5. *Supposons que l'observation X ait une loi $\mathbb{P}_\theta$ associé au modèle exponentiel suivant :*

$$\forall\ \theta \in \Theta, \forall\ x, \quad f(x,\theta) = h(x)\exp(a(\theta)T(x) - b(\theta)),$$

avec

- *$T(X)$ admet un moment d'ordre 2 par rapport à $\mathbb{P}_\theta$, $\forall\ \theta \in \Theta$,*
- *a et b 2 fois dérivables sur Θ tel que a' est non nul sur Θ et tel que $C(\theta) = \frac{b'(\theta)}{a'(\theta)}$ soit C^1 sur Θ de dérivée jamais nulle.*

Si $(X_1, \ldots, X_n)$ est un échantillon de même loi que X, si on cherche à estimer $g(\theta) = \theta$ et si $\hat{\theta}_n$ l'estimateur du maximum de vraisemblance construit à partir de cet échantillon existe, alors il vérifie :

$$\sqrt{n}(\hat{\theta}_n - \theta) \overset{n\to+\infty}{\longrightarrow} \mathcal{N}(0, 1/I(\theta)) \text{ en loi.}$$

Preuve : La log-vraisemblance associée au modèle est :

$$L_{X_1,\dots,X_n}(\theta) = \sum_{i=1}^{n} [\log h(X_i) + a(\theta)T(X_i) - b(\theta)].$$

Les hypothèses permettent d'affirmer que C^{-1} existe et d'écrire que

$$\hat{\theta}_n = C^{-1}\left(\overline{T}_n\right),$$

où $\overline{T}_n = \frac{1}{n}\sum_{i=1}^{n} T(X_i)$. On a

$$\sqrt{n}(\overline{T}_n - C(\theta)) \overset{n\to+\infty}{\longrightarrow} \mathcal{N}\left(0, \mathrm{var}_\theta(T(X))\right) \text{ en loi.}$$

Par le Théorème 1.1,

$$\sqrt{n}(\hat{\theta}_n - \theta) \overset{n\to+\infty}{\longrightarrow} \mathcal{N}\left(0, \frac{\mathrm{var}_\theta T(X)}{C'(\theta)^2}\right) \text{ en loi.}$$

La preuve du théorème précédent montre que

$$\frac{\mathrm{var}_\theta T(X)}{C'(\theta)^2} = \frac{1}{I(\theta)}.$$

□

Ce résultat est satisfaisant car si on est dans un cadre où la convergence en loi implique la convergence pour la norme L_2, et si $\hat{\theta}_n$ est sans biais, alors il est aussi asymptotiquement UMVU.

6.4 Conclusion

Dans les sections précédentes, on a fourni des réponses (qui restent partielles) à la conjecture de Fisher. Dans un cadre asymptotique, pour estimer $g(\theta)$, la meilleure limite possible en terme de distribution semble être la loi $\mathcal{N}(0, g'(\theta)^2/I(\theta))$. Mais la situation est plus compliquée en général et la notion d'optimalité reste, dans une large mesure, une question de goût. Il faut noter en particulier que les résultats précédents ont nécessité beaucoup d'hypothèses de régularité. En considérant des estimateurs peu réguliers (estimateurs par seuillage ou par contraction), on peut s'affranchir de la contrainte donnée par la borne de Cramer-Rao (estimation "super-efficace").

Chapitre 7

Calcul d'intégrales par méthodes de Monte Carlo

7.1 Introduction

Dans ce chapitre, nous nous intéressons en toute généralité au calcul numérique d'intégrales. Cette problématique sera envisagée d'un point de vue probabiliste et traitée en utilisant les méthodes de Monte Carlo. Nous illustrerons notre propos à l'aide des exemples suivants.

Exemple 7.1. L'expérience de Buffon (1777) pour calculer π.

Le lancer aléatoire et répété d'une aiguille de longueur l sur un plan strié de droites parallèles distantes les unes des autres de la longueur $d \geq l$. Si on note A l'événement "L'aiguille coupe une des droites du plan", on a

$$\mathbb{P}(A) = \mathbb{E}(1_A) = \frac{1}{\pi} \int_{-\pi/2}^{\pi/2} \frac{l \cos(x)}{d} dx = \frac{2l}{\pi d}.$$

Un calcul approché de $\mathbb{E}(1_A)$ fournit donc un calcul (ou plutôt une estimation) de π. Noter que ce ne sont pas des méthodes probabilistes qui donnent les meilleurs algorithmes pour calculer π. Ce sont des raisons historiques qui font que la méthode de Buffon est fréquemment citée.

Exemple 7.2. Alternatives pour calculer π.

Le calcul approché de π peut aussi s'effectuer en exploitant que les 3 intégrales suivantes ont chacune pour valeur π.

$$I_1 = \int_0^1 \frac{2}{\sqrt{1-x^2}} dx, \quad I_2 = \int_0^1 4\sqrt{1-x^2} dx, \quad I_3 = \int\int 1_{x^2+y^2 \leq 1} dxdy.$$

Exemple 7.3. Niveau réel d'un intervalle de confiance.

En utilisant le TCL ou les inégalités de Markov, de Bienaymé-Tchebychev, de Hoeffding ou de grandes déviations, on peut obtenir des intervalles de confiance de niveau α, mais on a rarement la valeur du niveau réel (on sait juste par construction que la probabilité d'être dans l'intervalle de confiance est supérieure à $1-\alpha$). Pour obtenir le niveau réel, comme précédemment, on est donc ramené au calcul d'une intégrale de la forme $\mathbb{E}(1_A)$.

Exemple 7.4. Un problème de finance.

On fait l'hypothèse qu'à un instant T (déterminé à l'avance), le prix d'une action est distribué comme $\exp(\beta_T X)$ où β_T est une constante et $X \sim \mathcal{N}(0,1)$ (modélisation de Black et Scholes). Un acheteur a la possibilité d'acheter à l'instant $t = T$ l'action au prix K fixé à l'instant $t = 0$ (option d'achat ou call). Le bénéfice qu'il peut espérer en revendant immédiatement l'action est donc

$$I = \mathbb{E}\left((\exp(\beta_T X) - K)_+\right),$$

quantité non calculable explicitement mais que l'on peut estimer. Cette quantité est appelée le prix du call. De la même manière, on définit l'option de vente (put) et le prix associé est

$$I = \mathbb{E}\left((K - \exp(\beta_T X))_+\right).$$

Dans ces quatre exemples, nous avons donc à calculer numériquement une intégrale I qui s'écrit sous la forme $I = \mathbb{E}(f(X))$ où f est explicite et X un vecteur aléatoire de $\mathbb{R}^d$. Il faut noter que toute intégrale peut s'écrire sous la forme d'une espérance d'un vecteur aléatoire. Dans la section suivante, nous décrivons la méthode de Monte Carlo pour calculer I et nous la comparons numériquement aux algorithmes déterministes. Nous utilisons sans les rappeler les méthodes de simulation de variables aléatoires (cf le cours de Sophie Lemaire).

7.2 Description et performance de la méthode de Monte Carlo

On cherche dans cette section à estimer $I = \mathbb{E}(f(X))$ où X est un vecteur aléatoire de $\mathbb{R}^d$. La méthode de Monte Carlo pour calculer I repose sur la loi forte des grands nombres :

Théorème 7.1. *Soit $X_1, \ldots, X_n$ un n-échantillon de même loi que X. Si $\mathbb{E}(|f(X)|) < \infty$, alors*

$$\frac{1}{n}\sum_{i=1}^{n} f(X_i) \overset{n\to+\infty}{\longrightarrow} I \quad p.s.$$

Pour calculer I, on simule donc $X_1, \dots, X_n$ n variables de même loi que X et on approche I par

$$I \approx \hat{I}_n = \frac{1}{n} \sum_{i=1}^{n} f(X_i).$$

La variance de l'erreur d'approximation de $\hat{I}_n$ est

$$v_n = \mathrm{var}(\hat{I}_n - I) = \frac{1}{n} \mathrm{var}(f(X)).$$

On peut aussi mesurer la qualité de cette approximation en utilisant le résultat suivant qui découle directement du TCL :

Théorème 7.2. *Soit $X_1, \dots, X_n$ un n-échantillon de même loi que X. Si $\mathbb{E}(f^2(X)) < \infty$, alors pour tous $a < b$,*

$$\mathbb{P}\left(\frac{a\sigma}{\sqrt{n}} \leq \hat{I}_n - I \leq \frac{b\sigma}{\sqrt{n}} \right) \overset{n \to +\infty}{\longrightarrow} \int_a^b \exp\left(-\frac{x^2}{2} \right) \frac{dx}{\sqrt{2\pi}},$$

avec $\sigma^2 = var(f(X))$. En particulier,

$$\mathbb{P}\left(|\hat{I}_n - I| \leq \frac{1.96\sigma}{\sqrt{n}} \right) \overset{n \to +\infty}{\longrightarrow} 0.95.$$

La précision de la méthode de Monte Carlo est donc au premier ordre en $O(n^{-1/2})$, ceci pour tout d et sous des hypothèses très faibles sur $f(X)$. Il est à noter qu'aucune hypothèse de régularité n'a été nécessaire. A titre de comparaison, la méthode des Trapèzes nécessite n^d mailles pour obtenir une précision en $O(n^{-2})$ et la méthode de Simpson nécessite n^d mailles pour obtenir une précision en $O(n^{-4})$. Par ailleurs, ces deux méthodes déterministes ne sont applicables que sous de fortes hypothèses de régularité. Pour une précision ε fixée, aux constantes près, la complexité de ces méthodes est donc :

MMC	Trapèzes	Simpson
ε^{-2}	$\varepsilon^{-d/2}$	$\varepsilon^{-d/4}$

Conclusion : La méthode de Monte Carlo est préférable quand la dimension est grande ou quand on n'a pas d'hypothèses de régularité.

Pour connaître précisément la précision (désolé !) de la méthode de Monte Carlo, il est nécessaire de connaître σ, ce qui est souvent au moins aussi difficile que de connaître I. Néanmoins, le lemme de Slutsky montre que le théorème précédent reste vrai en remplaçant σ par un estimateur consistant de cette quantité (par exemple en remplaçant σ par $s_n = ((n-1)^{-1} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2)^{1/2}$).

Enfin, il est important de noter que la méthode de Monte Carlo sera d'autant meilleure que la variance σ^2 est petite. Au passage, notons que la précision n'est pas contrôlable pour le calcul de I_1 dans le cas de l'Exemple 7.2. Dans la section suivante, on va donc s'attacher à décrire quelques méthodes pour réduire la variance.

7.3 Méthodes de réduction de la variance

On cherche toujours dans cette section à estimer $I = \mathbb{E}(f(X))$ où X est un vecteur aléatoire de $\mathbb{R}^d$. On va s'attacher à chercher des estimateurs basés sur la méthode de Monte Carlo construits à partir de n tirages indépendants de telle sorte que la variance de l'erreur d'approximation de chaque estimateur soit minimale. Remarquons que diminuer la variance ne signifie pas nécessairement obtenir un meilleur algorithme pour calculer I. C'est le nombre d'opérations élémentaires à effectuer pour atteindre une précision donnée qui compte.

7.3.1 Variable de contrôle

On veut estimer $I = \mathbb{E}(f(X))$ et on suppose qu'il existe g telle que $\mathbb{E}(g(X))$ est connue. L'idée est donc de poser

$$Y = f(X) + a(g(X)) - \mathbb{E}(g(X))),$$

où a est une constante. On a $\mathbb{E}(Y) = I$ et

$$\mathrm{var}(Y) = \mathrm{var}(f(X)) + a^2\mathrm{var}(g(X)) + 2a\mathrm{cov}(f(X), g(X)).$$

Donc

$$a \in \left]-\frac{2\mathrm{cov}(f(X), g(X))}{\mathrm{var}(g(X))}, 0\right[\Rightarrow \mathrm{var}(Y) < \mathrm{var}(f(X))$$

et $\mathrm{var}(Y)$ sera minimale pour

$$a = a_{f,g} = -\frac{\mathrm{cov}(f(X), g(X))}{\mathrm{var}(g(X))}.$$

Dans l'idéal, on choisit donc $a = a_{f,g}$ à condition que $a_{f,g}$ soit connue, ce qui est rarement le cas. Néanmoins, il est souvent possible de prendre a connue avec

$$a \in \left]-\frac{2\mathrm{cov}(f(X), g(X))}{\mathrm{var}(g(X))}, 0\right[$$

et d'appliquer la méthode de Monte Carlo à Y. Reprenons l'Exemple 7.4 en considérant le prix du call et du put :

$$\mathbb{E}\left((\exp(\beta_T X) - K)_+\right) - \mathbb{E}\left((K - \exp(\beta_T X))_+\right) = \mathbb{E}\left(\exp(\beta_T X) - K\right) = \exp\left(\frac{\beta_T^2}{2}\right) - K.$$

Il suffit d'estimer le prix d'une des options (call ou put) pour en déduire le prix de l'autre.

7.3.2 Variables corrélées négativement

Commençons par démontrer le résultat suivant :

Proposition 7.1. *Soit X un vecteur aléatoire de $\mathbb{R}^d$. Si f et g sont 2 fonctions à valeurs dans $\mathbb{R}$ croissantes en chacun de leurs arguments, on a :*

$$cov(f(X), g(X)) \geq 0.$$

Preuve : Traitons le cas $d = 1$. On utilise que

$$\forall\, (x, y) \in \mathbb{R}^2, \quad (f(x) - f(y))(g(x) - g(y)) \geq 0.$$

Donc

$$\mathbb{E}((f(X) - f(Y))(g(X) - g(Y))) \geq 0.$$

Avec $X \sim Y$ et X et Y indépendantes, on conclut pour le cas $d = 1$. Le cas $d > 1$ s'obtient par récurrence et en conditionnant. Si $X = (X_1, \ldots, X_d)$ on a par hypothèse de récurrence au rang $d - 1$:

$$\mathbb{E}(f(X)g(X)|X_d) \geq \mathbb{E}(f(X)|X_d)\mathbb{E}(g(X)|X_d),$$

on applique le cas $d = 1$ pour minorer l'espérance du terme de droite. □

A présent, supposons que l'on veuille calculer $I = \mathbb{E}(f(X))$ de telle sorte qu'il existe $h : \mathbb{R}^d \longrightarrow \mathbb{R}^d$ une fonction décroissante en chacun de ses arguments telle que $h(X) \sim X$. Alors, on pose

$$Y = \frac{1}{2}(f(X) + f(h(X))).$$

On a :

$$\mathbb{E}(Y) = \mathbb{E}(f(X)), \quad \text{var}(Y) = \frac{1}{2}(\text{var}(f(X)) + \text{cov}(f(X), f(h(X))).$$

Donc si f est une fonction monotone en chacun de ses arguments,

$$\text{cov}(f(X), f(h(X)) \leq 0.$$

Conclusion : Sous les hypothèses précédentes de monotonie portant sur f et h, l'application de la méthode de Monte Carlo sur un n-échantillon de même loi que celle de Y sera préférable à l'application de la méthode de Monte Carlo sur un $2n$-échantillon de même loi que celle de $f(X)$. Cette méthode s'applique facilement aux Exemples 7.2 et 7.4.

7.3.3 Echantillonnage préférentiel

On cherche toujours à calculer $I = \mathbb{E}(f(X))$ et on suppose que X admet g pour densité par rapport à la mesure de Lebesgue. On a avec $\tilde{g}$ densité strictement positive :

$$\begin{aligned} I &= \int f(x)g(x)dx \\ &= \int \frac{f(x)g(x)}{\tilde{g}(x)}\tilde{g}(x)dx \\ &= \mathbb{E}\left(\frac{f(Y)g(Y)}{\tilde{g}(Y)}\right) \end{aligned}$$

où Y a pour densité $\tilde{g}$. On peut donc appliquer la méthode de Monte Carlo sur un n-échantillon de même loi que celle de $Z = \frac{f(Y)g(Y)}{\tilde{g}(Y)}$. Cela n'aura un intérêt que si $\tilde{g}$ est choisie de telle sorte que

$$\mathrm{var}(Z) = \int \frac{g^2(x)f^2(x)}{\tilde{g}(x)}dx - (\mathbb{E}(f(X)))^2$$

soit inférieure à $\mathrm{var}(f(X))$. On vérifie que le choix idéal est

$$\tilde{g}(x) = \frac{f(x)g(x)}{\mathbb{E}(f(X))}$$

pour lequel $\mathrm{var}(Z) = 0$. Ce choix est bien entendu impossible car il nécessite la connaissance de $\mathbb{E}(f(X))$. Néanmoins, il donne des idées.

- Reprenons l'Exemple 7.2 avec I_2. Un choix naturel est de prendre $\tilde{g}(x) = \frac{6}{5}\left(1 - \frac{x^2}{2}\right)1_{x\in[0,1]}$.
- Considérons le put de l'Exemple 7.4 avec $K = 1$. En observant que $e^x - 1 \approx x$ au voisinage de 0, on écrit :

$$\begin{aligned} \mathbb{E}\left((1 - \exp(\beta_T X))_+\right) &= \int_{-\infty}^{+\infty} (1 - e^{\beta_T x})_+ \frac{e^{-x^2/2}}{\sqrt{2\pi}}dx \\ &= \int_{-\infty}^{+\infty} \frac{(1 - e^{\beta_T x})_+}{\beta_T|x|}\beta_T|x|\frac{e^{-x^2/2}}{\sqrt{2\pi}}dx \\ &= \mathbb{E}\left(\frac{(1 - \exp(\beta_T\sqrt{Y}))_+ + (1 - \exp(-\beta_T\sqrt{Y}))_+}{\sqrt{2\pi}\sqrt{Y}}\right), \end{aligned}$$

où $Y \sim \mathcal{E}(1/2)$.

La mise en oeuvre de cette approche montre une amélioration sensible de la précision du calcul pour ces deux exemples.

7.3.4 Conditionnement

Commençons par démontrer le résultat suivant :

Proposition 7.2. *Pour tous vecteurs aléatoires U et V de carré intégrable, en notant $Z = \mathbb{E}(U|V)$, on a :*

$$var(U) = \mathbb{E}(U - Z)^2 + var(Z).$$

Preuve : immédiate en calculant $\mathbb{E}(U - Z)^2$. □

Le conditionnement permet donc de diminuer la variance quand on sait simuler la loi de $\mathbb{E}(f(X)|Y)$, avec Y donné.

7.3.5 Stratification

On cherche toujours à estimer à l'aide de n tirages $I = \mathbb{E}(f(X))$ où X est un vecteur aléatoire de $\mathbb{R}^d$ de densité g par rapport à la mesure de Lebesgue. On se donne une partition $(D_i, 1 \leq i \leq m)$ de $\mathbb{R}^d$ où $m \in \mathbb{N}^*$. On décompose alors I de la façon suivante :

$$I = \sum_{i=1}^{m} \mathbb{E}(1_{X \in D_i} f(X)) = \sum_{i=1}^{m} \mathbb{E}(f(X)|X \in D_i) p_i$$

où $p_i = \mathbb{P}(X \in D_i)$. Lorsque l'on connaît les nombres p_i, on peut utiliser une méthode de Monte Carlo pour estimer les intégrales $I_i = \mathbb{E}(f(X)|X \in D_i)$. On obtient ainsi m estimateurs $\tilde{I}_i$ obtenus chacun avec n_i tirages indépendants tels que $\sum_{i=1}^{m} n_i = n$. La variance de l'erreur d'approximation de chaque estimateur est donnée par

$$\forall i \in \{1, \ldots, m\}, \quad \frac{1}{n_i}\sigma_i^2 = \frac{1}{n_i}\mathrm{var}(f(X)|X \in D_i).$$

L'estimateur de I est

$$\sum_{i=1}^{m} p_i \tilde{I}_i$$

de variance

$$\tilde{v}_n = \sum_{i=1}^{m} p_i^2 \frac{\sigma_i^2}{n_i}.$$

Il reste alors à choisir les n_i de manière à minimiser $\tilde{v}_n$ (on ne s'intéressera pas au choix de la partition). Le calcul montre que l'on doit choisir

$$n_i = \frac{n p_i \sigma_i}{\sum_{i=1}^{m} p_i \sigma_i}$$

et on obtient ainsi

$$\tilde{v}_n = \frac{1}{n}\left(\sum_{i=1}^{m} p_i \sigma_i\right)^2.$$

On a alors :

$$\begin{aligned}
v_n &= \frac{1}{n}\left(\mathbb{E}(f^2(X)) - (\mathbb{E}(f(X)))^2\right) \\
&= \frac{1}{n}\left(\sum_{i=1}^{m} p_i \mathbb{E}(f^2(X)|X \in D_i) - \left(\sum_{i=}^{m} p_i \mathbb{E}(f(X)|X \in D_i)\right)^2\right) \\
&= \frac{1}{n}\left(\sum_{i=1}^{m} p_i \sigma_i^2 + \sum_{i=1}^{m} p_i I_i^2 - (\sum_{i=1}^{m} p_i I_i)^2\right) \\
&\geq \frac{1}{n}\left(\sum_{i=1}^{m} p_i \sigma_i^2\right) \\
&\geq \frac{1}{n}\left(\sum_{i=1}^{m} p_i \sigma_i\right)^2 \\
&= \tilde{v}_n.
\end{aligned}$$

Ceci prouve que par stratification, on peut obtenir un estimateur de variance moindre. Le problème est qu'il faut connaître les p_i et les σ_i que l'on peut estimer en première approche. Néanmoins il existe une stratégie plus souple qui consiste à choisir $\forall\ i \in \{1, \ldots, m\}$, $n_i = np_i$. On obtient alors un estimateur de variance

$$\frac{1}{n}\left(\sum_{i=1}^{m} p_i \sigma_i^2\right).$$

Pour illustrer simplement cette méthode, considérons le cas du call dans l'Exemple 7.4. On considère naturellement $m = 2$, et la partition telle que $D_1 = \{x > \log(K)/\beta_T\}$ et $D_2 = \{x \leq \log(K)/\beta_T\}$. Dans ce cas là, $\sigma_2 = 0$ et on affecte les n points dans I_1.

Bibliographie

Azaïs, J.M. et Bardet J.M. (2005) *Le modèle linéaire par l'exemple.* Dunod.

Bickel, P.J. et Doksum, K.A. (2000) *Mathematical Statistics : Basic Ideas and Selected Topics.* Prentice Hall.

Billingsley, P. (1986) *Probability and measure. Second edition.* John Wiley & Sons, Inc., New York.

Bon, J.L (1986) *Fiabilité des systèmes : Méthodes mathématiques.* Dunod.

Cottrell, M., Genon-Catalot, V., Duhamel C. et Meyre, T. (2005) *Exercices de probablités. Licence, master, écoles d'ingénieurs.* Cassini.

Dacunha-Castelle, D. et Duflo, M. (1997) *Probabilités et statistiques.* Masson.

Daudin, J.J., Robin. S. et Vuillet, C. (2001) *Statistique inférentielle. Idées, démarches, exemples.* Société française de statistique.

Fourdriner, D. (2002) *Statistique inférentielle.* Dunod.

Genon-Catalot, V. et Picard, D. (1993) *Eléments de statistique asymptotique.* Springer.

Lapeyre, B. Pardoux, E. et Sentis, R. (1998) *Méthodes de Monte Carlo pour les équations de transport et de diffusion.* Springer-Verlag.

Milhaud, X. (1991) *Statistique.* Belin.

Monfort, A. (1997) *Cours de statistique mathématique.* Economica

Ouvrard, J.Y. (2000) *Probabilités 2. Maîtrise agrégation.* Cassini.

Picard, D. (1997) *Cours de statistique.* http ://www.proba.jussieu.fr/pageperso/picard/picard.html

Prum, B. (1996) *Modèle linéaire. Comparaison de groupes et régression.* INSERM.

Revuz, D. (1987) *Probabilités.* Hermann.

Saporta, G. (1990) *Probabilités, analyse des données et statistique.* Technip.

Tsybakov, A.B. (1990) *Introduction à l'estimation non-paramétrique*. Springer.

Van der Vaart, A.W. (2000) *Asymptotic Statistics*. Cambridge University Press.

Wasserman, L. (2005) *All of statistics. A concise course in statistical inference*. Springer.

Table des matières