Cours de Statistique de Base

Anne PHILIPPE[1] & Marie-Claude VIANO[2]

[1] Université de Nantes
Département de Mathématiques
2 rue de la Houssinière - BP 92208 -
F-44322 Nantes Cedex 3

[2] Université de Lille 1

Table des matières

Table des figures

Liste des tableaux

CHAPITRE 1

Introduction

Le statisticien se voit confronté à des données.

EXEMPLE 1.1. le taux de cholestérol mesuré sur 200 femmes âgées de 50 ans. ‖

EXEMPLE 1.2. le taux d'une certaine hormone mesuré sur 40 personnes d'abord avant traitement médical, puis après traitement médical. ‖

De ces données il va tirer une conclusion sur la population globale. Cette démarche s'appelle l'*inférence statistique*.

Dans l'exemple 1.1, il pourra vouloir se faire une idée du taux moyen de cholestérol de la population des femmes âgées de 50 ans. C'est un problème d'*estimation*. Dans l'exemple 1.2 il lui sera sans doute demandé si le traitement médical est efficace ou non. C'est un problème de *test*.

Dans l'exemple 1.1, notons $x_1, \cdots, x_{200}$ les données. Le premier réflexe consiste à en calculer la moyenne arithmétique $\overline{x} = \frac{x_1+\cdots+x_{200}}{200}$. Que faire ensuite de la valeur obtenue ? En quoi représente t'elle le taux moyen de cholestérol de la population des femmes de 50 ans ? Peut-on donner un encadrement (on parle aussi de *fourchette* ou d'*intervalle de confiance*) de ce taux moyen inconnu ? Si oui, quelle confiance accorder à cet encadrement ?

Dans l'exemple 1.2, notons $(u_1, v_1), \cdots, (u_{40}, v_{40})$ les données. On calcule les deux moyennes arithmétiques $\overline{u}$ et $\overline{v}$ et on les compare. Admettons qu'on ait obtenu $\overline{u} = 1.8$ et $\overline{v} = 1.9$. Peut-on en déduire que le médicament est efficace pour augmenter le taux de l'hormone considérée ? Ou vaut-il mieux penser que le changement de 1.8 à 1.9 est tout simplement dû aux erreurs de mesure et aux fluctuations normales au cours du temps du taux d'hormone chez chaque personne ?

D'une manière générale, il est évident qu'on ne peut donner une conclusion dépassant la simple description des données qu'au prix *d'hypothèses probabilistes*. Reprenons l'exemple 1.2.
On traite les données comme les réalisations de vecteurs aléatoires $(U_1, V_1), \cdots, (U_{40}, V_{40})$ et on fait des hypothèses sur la loi de ces vecteurs (par exemple : ils sont indépendants entre eux, ils ont même loi et cette loi est gaussienne). Ces hypothèses permettent d'obtenir une échelle de comparaison qui nous dira si on peut juger que 1.8 est *significativement* plus petit que 1.9.

Toute procédure d'inférence statistique est basée sur de telles hypothèses, (que le praticien, souvent, a oubliées...). Le but de ce cours est de donner à la fois quelques

procédures courantes et leurs fondements théoriques.

Le schéma théorique de la plupart des problèmes de statistique inférentielle est le suivant.
1- Un ensemble mesurable $(\Omega, \mathcal{A})$.
2- Cet espace est muni non pas d'une probabilité mais de toute une famille $(\Pi_\theta)_{\theta\in\Theta}$. Θ sera dans ce cours un ouvert de $\mathbb{R}^p$ et θ est le paramètre, inconnu, qu'il s'agit d'estimer ou sur lequel il s'agit de tirer une conclusion quant à sa localisation dans Θ. Dans l'exemple 1.2, θ est le couple (μ_1, μ_2) des espérances de la loi de U_1 et de celle de V_1, et il s'agit de tester si $\mu_1 - \mu_2$ est négatif.
3- Sur l'espace $(\Omega, \mathcal{A})$, on se donne n variables aléatoires toutes de même loi, $(X_1, \cdots, X_n)$, à valeurs dans un espace mesurable $(\underline{X}, \mathcal{X})$ ($\underline{X} = \mathbb{R}$ dans l'exemple 1.1, ou $\underline{X} = \mathbb{R}^2$ dans l'exemple 1.2, $\underline{X} = \mathbb{N}^*$ dans l'exemple 2.1 ci-dessous etc...). Les données dont dispose le praticien sont les valeurs $(X_1(\omega), \cdots, X_n(\omega))$ que prennent ces variables aléatoires pour un certain élément $\omega \in \Omega$.
4- On note P_θ l'image de Π_θ par l'une quelconque des variables X_j
5- Dans les exemples élémentaires (ce sera le cas le plus souvent dans ce cours), les X_j sont supposées indépendantes. Dans cette situation standard, $(X_1, \cdots, X_n)$ est appelé n-échantillon de la loi P_θ. Le nombre n est appelé taille de l'échantillon. Dans le cas de variables indépendantes, l'image de la famille $(\Omega, \mathcal{A}, (\Pi_\theta)_{\theta\in\Theta})$ par le n-échantillon est, bien entendu, la famille d'espaces produits

$$(\underline{X}^n, \mathcal{X}^{\otimes n}, (P_\theta^{\otimes n})_{\theta\in\Theta}).$$

Comme toujours en Probabilités, on peut oublier la structure $(\Omega, \mathcal{A}, (\Pi_\theta)_{\theta\in\Theta})$ de départ et ne travailler qu'avec les espaces images, donc, dans le cas de variables indépendantes, avec cette famille d'espaces produits.
Une des difficultés principales de ce cours vient du fait qu'il n'y a pas une seule probabilité mais toute une famille. Les notations ont de ce fait une importance décisive. On notera $\mathbb{E}_\theta$, Var_θ, etc... l'espérance, la variance, etc... calculées pour la valeur θ du paramètre, c'est à dire lorsque la loi des X_j est P_θ. De la même façon, le cas échéant, on notera $\mathbb{E}_\mu$, Var_μ, etc... les moments lorsque la loi est μ.

Revu dans le cadre de ce schéma théorique, le travail du statisticien consiste à s'appuyer sur les variables aléatoires $(X_1, \cdots, X_n)$ pour localiser dans Θ le paramètre inconnu θ.

CHAPITRE 2

Estimation ponctuelle

Soit $(X_1, \cdots, X_n)$ un $n-$échantillon (*i.e.* n variables aléatoires i.i.d.) suivant la loi P_θ avec $\theta \in \Theta$ un ouvert de $\mathbb{R}^p$.

Soit g est une application mesurable de Θ dans $\mathbb{R}^k$. Il s'agit ici de proposer un **estimateur** de $g(\theta)$, c'est à dire une fonction mesurable de $(X_1, \cdots, X_n)$ qui donne "une idée satisfaisante" de $g(\theta)$. On notera $\hat{g}_n(X_1, \cdots, X_n)$ cet estimateur.

1. Définitions et exemples

DÉFINITION 2.1. *On dit que $\hat{g}_n(X_1, \cdots, X_n)$ est un estimateur sans biais de $g(\theta)$ si*

$$\mathbb{E}_\theta(\hat{g}_n(X_1, \cdots, X_n)) = g(\theta) \quad \forall \theta \in \Theta.$$

Évidemment, on appelle biais de l'estimateur la différence $\mathbb{E}_\theta(\hat{g}_n(X_1, \cdots, X_n)) - g(\theta)$ qui sépare son espérance de la quantité qu'il est censé estimer. Si le biais tend vers zéro lorsque $n \to \infty$, on dit que l'estimateur est *asymptotiquement sans biais.*

DÉFINITION 2.2. *On dit que l'estimateur est convergent (en moyenne quadratique, presque sûrement, en probabilité etc...) si, lorsque la valeur du paramètre est θ, la suite de variables aléatoires $\hat{g}_n(X_1, \cdots, X_n)$ converge (en moyenne quadratique, presque sûrement, en probabilité etc...) vers $g(\theta)$ quand $n \to \infty$.*

Parce qu'il est souvent facile de calculer les moments d'ordre 1 et 2 (lorsqu'ils existent !), on attache une particulière importance à la convergence en moyenne quadratique. L'erreur quadratique

$$\mathbb{E}_\theta \left\| \hat{g}_n(X_1, \cdots, X_n) - g(\theta) \right\|_2^2$$

facture l'erreur d'estimation (dans cette expression, $\|u\|_2^2 = u_1^2 + \cdots + u_k^2$ désigne la norme quadratique habituelle de $\mathbb{R}^k$). Lorsque l'estimateur est sans biais et à valeurs dans $\mathbb{R}$, cette erreur quadratique est tout simplement la variance de l'estimateur.
L'erreur quadratique tend vers zéro si et seulement si l'estimateur converge en moyenne quadratique.

EXEMPLE 2.1. P_θ est la loi de Poisson de paramètre θ. Ici, $\Theta =]0, +\infty[$. A partir du n-échantillon $(X_1, \cdots, X_n)$ de variables aléatoires à valeurs entières la première idée est d'estimer θ par $\overline{X}_n = \frac{X_1 + \cdots + X_n}{n}$. Cet estimateur est sans biais et convergent en moyenne quadratique et presque sûrement. ‖

EXEMPLE 2.2. P_θ est la loi uniforme sur $[0, \theta]$. Ici encore $\Theta =]0, +\infty[$. A partir du n-échantillon $(X_1, \cdots, X_n)$ de variables à valeurs dans $\mathbb{R}^+$ on peut estimer le paramètre θ par $\max\{X_1, \cdots, X_n\}$. Cet estimateur est asymptotiquement sans

biais. Il est convergent en moyenne quadratique et presque sûrement.
‖

EXEMPLE 2.3. P_θ est la loi de Gauss de paramètre inconnu $\theta = (m, \sigma^2)$. Ici, $\Theta = \mathbb{R}\times]0, +\infty[$. A partir du n-échantillon $(X_1, \cdots, X_n)$ de variables à valeurs dans $X = \mathbb{R}$ on peut estimer m par $\overline{X}_n$ et σ^2 par $\frac{(X_1-\overline{X}_n)^2+\cdots+(X_n-\overline{X}_n)^2}{n-1}$. Quelles sont les propriétés de ces estimateurs ?
Il faut remarquer que, dans le contexte présent où ni l'espérance ni la variance ne sont supposées connues, la variable aléatoire $\frac{(X_1-m)^2+\cdots+(X_n-m)^2}{n}$ *n'est pas* un estimateur de la variance, tout simplement parce que sa valeur sur l'échantillon n'est pas calculable par le statisticien qui ne connaît pas m. ‖

On introduit parfois, entre estimateurs, la relation de pré-ordre

$$\hat{\theta}_1 \gg \hat{\theta}_2 \iff \mathbb{E}_\theta \|\hat{\theta}_1 - \theta\|_2^{\,2} \leq \mathbb{E}_\theta \|\hat{\theta}_2 - \theta\|_2^{\,2} \quad \forall \theta.$$

C'est bien sûr un pré-ordre partiel. On dira que $\hat{\theta}_1$ est "strictement meilleur" que $\hat{\theta}_2$ si $\hat{\theta}_1 \gg \hat{\theta}_2$ et si l'inégalité entre les erreurs quadratiques est stricte pour au moins une valeur de θ.

DÉFINITION 2.3. *Admissibilité*
L'estimateur $\hat{\theta}$ est admissible si il n'en existe pas de meilleur. Autrement dit si il est un élément maximal pour le pré-ordre partiel qu'on vient d'introduire.

2. Quelques outils théoriques et techniques

DÉFINITION 2.4. *On dit que le modèle $(\underline{X}^n, \mathcal{X}^{\otimes n}, (P_\theta^{\otimes n})_{\theta\in\Theta})$ est dominé si il existe une mesure σ-finie μ qui domine toutes les probabilités $(P_\theta)_{\theta\in\Theta}$.*

Ceci signifie que chaque P_θ a une densité par rapport à μ. En d'autres termes, le modèle est dominé si il existe une famille d'applications mesurables de $\underline{X}$ dans $\mathbb{R}^+$, $(f(x,\theta))_{\theta\in\Theta}$ telles que pour toute variable aléatoire intégrable U sur l'espace $\underline{X}$

$$\mathbb{E}_\theta(U(X)) = \mathbb{E}_\mu\left(f(X,\theta)U(X)\right). \tag{2.1}$$

Puisque les X_j sont indépendantes, ceci revient à dire que, pour toute fonction intégrable de l'échantillon $U(X_1, \cdots, X_n)$,

$$\mathbb{E}_\theta(U(X_1, \cdots, X_n)) = \mathbb{E}_\mu\left(U(X_1, \cdots, X_n)\prod_{j=1}^n f(X_j, \theta)\right). \tag{2.2}$$

où $\mathbb{E}_\mu$ et $\mathbb{E}_\theta$ désignent (abusivement !) l'intégration par rapport à $\mu^{\otimes n}$ et à $P_\theta^{\otimes n}$.

REMARQUE 2.1. Lorsque les variables ne sont pas indépendantes, la domination des P_θ n'implique pas celle des lois de $(X_1, \cdots, X_n)$. On suppose directement qu'il existe une mesure ν sur $(\underline{X}^n, \mathcal{X}^{\otimes n})$ telle que, pour tout θ, la loi de $(X_1, \cdots, X_n)$ sous Π_θ est dominée par ν. Mais cette situation ne sera rencontrée que dans quelques exercices. |

Reprenons les trois exemples.

– Dans l'exemple 2.1, la mesure dominante est la mesure dénombrement et la densité est

$$f(x,\theta) = e^{-\theta}\frac{\theta^x}{x!} \quad x = 0, 1, \cdots$$

– Dans l'exemple 2.2 c'est la mesure de Lebesgue sur $\mathbb{R}^+$ et la densité est

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

– Dans l'exemple 2.3, la mesure dominante est la mesure de Lebesgue sur $\mathbb{R}^2$ et la densité est celle, bien connue, de la loi gaussienne de paramètres (m,θ).

Quand le modèle est dominé par la mesure μ, on peut toujours supposer (et c'est ce qu'on fera par endroits) que cette mesure est une probabilité.

En effet comme la mesure μ qui domine la famille des P_θ est supposée $\sigma - finie$, il existe une partition mesurable $(A_j)_{j\geq 1}$ de $\underline{X}$ telle que $0 < \mu(A_j) < +\infty, \quad \forall j$. On définit la mesure $\mu^\star$ par

$$\mu^\star(A) = \sum_{j=1}^{\infty} \frac{1}{2^j} \frac{\mu(A \cap A_j)}{\mu(A_j)}.$$

C'est une probabilité et elle domine aussi la famille des P_θ.

DÉFINITION 2.5. *Supposons que le modèle est dominé et que les variables X_j sont indépendantes. On appelle vraisemblance de l'échantillon la densité de $P_\theta^{\otimes n}$ par rapport à $\mu^{\otimes n}$, soit*

$$V_\theta(x_1,\cdots,x_n) := \prod_{j=1}^{n} f(x_j,\theta) \quad x_j \in \underline{X}\ \forall j.$$

Pour les exemples précédents, la vraisemblance s'écrit

– Dans l'exemple 2.1,

$$V_\theta(x_1,\cdots,x_n) = e^{-n\theta}\frac{\theta^{\sum_{j=1}^n x_j}}{\prod_{j=1}^n x_j!}.$$

– Dans l'exemple 2.2,

$$V_\theta(x_1,\cdots,x_n) = \frac{1}{\theta^n}\mathbb{I}_{[0,\theta]}\left(\max\{x_1,\cdots,x_n\}\right),$$

– Dans l'exemple 2.3,

$$V_\theta(x_1,\cdots,x_n) = \frac{1}{(\sigma\sqrt{2\pi})^n} e^{-\frac{\sum_{j=1}^n (x_j-m)^2}{2\sigma^2}}.$$

Dans le cas où les densités $f(x,\theta)$ sont strictement positives quels que soient x et θ (plus exactement dans le cas où, pour tout θ il existe une version partout non nulle de la densité de P_θ par rapport à μ, ce qui revient à dire que la mesure dominante μ est dominée par toutes les mesures P_θ, les densités correspondantes étant alors $f^{-1}(x,\theta)$) il est souvent commode de travailler avec le logarithme de la vraisemblance

$$L_\theta(x_1,\cdots,x_n) := \ln V_\theta(x_1,\cdots,x_n) = \sum_{j=1}^{n} \ln f(x_j,\theta).$$

Les deux définitions qui suivent sont écrites dans le cas où le paramètre est à valeurs réelles : autrement dit Θ est un ouvert de $\mathbb{R}$

DÉFINITION 2.6. *On dit que le modèle $(\underline{X}^n, \mathcal{X}^{\otimes n}, (P_\theta^{\otimes n})_{\theta\in\Theta})$ est régulier si il est dominé et si*

(1) $f(x,\theta) > 0$ pour tout (x,θ),

(2) *pour tout x, la fonction $f(x,\theta)$ est dérivable en tout point de Θ par rapport à θ*

(3) *pour tout $\theta \in \Theta$,*

$$\int_{\underline{X}} \left| \frac{\partial f(x,\theta)}{\partial \theta} \right| d\mu(x) < \infty \text{ et } \int_{\underline{X}} \frac{\partial f(x,\theta)}{\partial \theta} d\mu(x) = 0 \tag{2.3}$$

(4) *Pour tout $\theta \in \Theta$,*

$$I(\theta) := \mathbb{E}_\theta \left(\frac{\partial \ln f(X,\theta)}{\partial \theta} \right)^2 > 0. \tag{2.4}$$

Dans les exemples précédents :
– Dans l'exemple 2.1, le modèle est régulier
– Dans l'exemple 2.2, il ne l'est pas.
Remarques avant de continuer :

(1) La condition (2.3) est la dérivabilité sous le signe intégrale. Il suffit de remarquer que l'intégrale $\int_{\underline{X}} f(x,\theta) d\mu(x) \equiv 1$,est partout dérivable sur Θ.

(2) La quantité $I(\theta)$ définie en (2.4), s'appelle *information de Fisher*. En utilisant la relation (2.1), cette information peut s'écrire aussi sous la forme

$$I(\theta) = \mathbb{E}_\mu \left(\frac{\left(\frac{\partial f(X,\theta)}{\partial \theta} \right)^2}{f(X,\theta)} \right)$$

(3) Une conséquence immédiate de (2.3) est que

$$\mathbb{E}_\theta \left(\frac{\frac{\partial f(X,\theta)}{\partial \theta}}{f(X,\theta)} \right) = \mathbb{E}_\theta \left(\frac{\partial \ln f(X,\theta)}{\partial \theta} \right) = \int_{\underline{X}} \frac{\partial f(x,\theta)}{\partial \theta} d\mu(x) \equiv 0, \tag{2.5}$$

ce qui signifie que la variable aléatoire $\frac{\partial ln f(X,\theta)}{\partial \theta}$ est centrée sous la loi P_θ.

(4) Revenant alors à l'information de Fisher, on voit que c'est tout simplement la variance de $\frac{\partial \ln f(X,\theta)}{\partial \theta}$

(5) De ce fait, la condition (2.4) qui affirme la stricte positivité de l'information de Fisher dit simplement que $f(X,\theta)$ n'est pas (presque sûrement) constante lorsque θ varie. On voit bien là en quoi cette condition affirme que la variable aléatoire X apporte une information sur le paramètre θ.

(6) Au prix d'hypothèses de régularité supplémentaires l'information de Fisher s'écrit aussi

$$I(\theta) = -\mathbb{E}_\theta \left(\frac{\partial^2 \ln f(X,\theta)}{\partial \theta^2} \right). \tag{2.6}$$

Ces conditions sont : $I(\theta) < \infty$ et une condition de dérivabilité deux fois sous signe intégrale, soit

$$\int_{\underline{X}} |\frac{\partial^2 f(x,\theta)}{\partial \theta^2}| d\mu(x) < \infty \quad \text{et} \int_{\underline{X}} \frac{\partial^2 f(x,\theta)}{\partial \theta^2} d\mu(x) = \mathbb{E}_\mu \left(\frac{\partial^2 f(X,\theta)}{\partial \theta^2} \right) \equiv 0. \tag{2.7}$$

DÉMONSTRATION. Sous ces conditions on a

$$\frac{\partial^2 \ln f(x,\theta)}{\partial\theta^2} = \frac{\partial^2 f(x,\theta)}{\partial\theta^2} f^{-1}(x,\theta) - \left(\frac{\partial f(x,\theta)}{\partial\theta} f^{-1}(x,\theta)\right)^2.$$

D'où

$$\mathbb{E}_\theta\left(\frac{\partial^2 \ln f(X,\theta)}{\partial\theta^2}\right) = \mathbb{E}_\mu\left(\frac{\partial^2 f(X,\theta)}{\partial\theta^2}\right) - I(\theta).$$

En utilisant (2.7) la preuve de (2.6) est terminée □

(7) Si la mesure dominante μ est remplacée par une mesure $\mu^\star$ qui lui est équivalente alors $f(x,\theta)$ devient $f(x,\theta)\phi(x)$ où ϕ est la densité (strictement positive) de μ par rapport à $\mu^\star$, et rien n'est changé dans le calcul de la dérivée du logarithme. L'information de Fisher ne change pas.

DÉFINITION 2.7. *Toujours dans le cas d'un modèle régulier, l'information de Fisher apportée par l'échantillon est la quantité*

$$I_n(\theta) := \mathbb{E}_\theta\left(\frac{\partial L_\theta(X_1,\cdots,X_n)}{\partial\theta}\right)^2.$$

On remarque que, si les variables sont indépendantes, la forme additive de L_θ permet d'écrire que

$$I_n(\theta) = \mathbb{E}_\theta\left(\sum_{j=1}^n \frac{\partial \ln f(X_j,\theta)}{\partial\theta}\right)^2 = \mathrm{Var}_\theta\left(\sum_{j=1}^n \frac{\partial \ln f(X_j,\theta)}{\partial\theta}\right) = nI(\theta),$$

donc, l'information apportée par un n-échantillon est n fois celle apportée par une seule variable. Ceci vaut évidemment si l'information est finie. Dans tout ce qui précède, on ne l'a pas supposé.
Dans l'exemple 2.1, il est facile de vérifier que

$$I_n(\theta) = n\mathrm{Var}_\theta\left(\frac{X}{\theta} - 1\right) = \frac{n}{\theta}.$$

Bien sûr ceci ne s'applique pas à l'exemple 2.3 où le paramètre a deux composantes. Lorsque θ est à valeurs dans $\mathbb{R}^p$ ($p \geq 2$), les définitions de l'information de Fisher sont à adapter. On définit alors une matrice d'information. Notons $\theta = {}^t(\theta_1,\cdots,\theta_p)$ le paramètre. Pour une fonction ϕ définie sur $\mathbb{R}^p$, soit

$$\nabla_\theta\phi = (\frac{\partial\phi}{\partial\theta_1},\cdots,\frac{\partial\phi}{\partial\theta_p})$$

son gradient.

DÉFINITION 2.8. *On dit que le modèle est régulier si il vérifie, outre les conditions 1 et 2 de la définition 2.6,*

(3 bis) Pour tout $\theta \in \Theta$

$$\int_{\underline{X}} \left|\frac{\partial f(x,\theta)}{\partial\theta_j}\right| d\mu(x) < \infty \text{ et } \int_{\underline{X}} \frac{\partial f(x,\theta)}{\partial\theta_j} d\mu(x) \equiv 0 \quad \forall\theta_j \tag{2.8}$$

(4 bis) pour tout θ, la matrice

$$I(\theta) := \mathbb{E}_\theta\left({}^t\nabla_\theta \ln f(X,\theta) \nabla_\theta \ln f(X,\theta)\right) \tag{2.9}$$

est non dégénérée.

On définit de la même façon $I_n(\theta)$, information apportée par l'échantillon. Dans le cas où les variables sont indépendantes on a encore

$$I_n(\theta) = nI(\theta).$$

Il est aisé de vérifier que, dans l'exemple 2.3, le modèle est régulier. On trouve

$$\begin{aligned} I_n(\theta) &= \begin{pmatrix} \mathbb{E}_\theta\left(\frac{\sum_{j=1}^n (X_j-m)}{\sigma^2}\right)^2 & \mathbb{E}_\theta\left(\frac{-n\sum_{j=1}^n (X_j-m)}{2\sigma^4} + \frac{\sum_{j=1}^n (X_j-m)^3}{2\sigma^4}\right) \\ \mathbb{E}_\theta\left(\frac{-n\sum_{j=1}^n (X_j-m)}{2\sigma^4} + \frac{\sum_{j=1}^n (X_j-m)^3}{2\sigma^4}\right) & \mathbb{E}_\theta\left(\frac{-n}{2\sigma^2} + \frac{\sum_{j=1}^n (X_j-m)^2}{2\sigma^4}\right)^2 \end{pmatrix} \\ &= \begin{pmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{pmatrix} \end{aligned}$$

3. Inégalité de Cramér-Rao

Dans le cas d'un modèle régulier et lorsque l'information de Fisher est finie, on connaît une borne inférieure pour la variance des estimateurs sans biais. C'est intéressant, car si on dispose d'un estimateur de L^2, sans biais, dont la variance est égale à cette borne, on sait qu'il est meilleur que tous les autres estimateurs sans biais (au sens de l'erreur quadratique en tous cas !).
Commençons par le cas où le paramètre est uni-dimensionnel. Supposons que le modèle est dominé. Soit $\hat{g}_n = \hat{g}_n(X_1, \cdots, X_n)$ un estimateur sans biais de $g(\theta)$. Supposons que g est dérivable sur Θ et que (dérivabilité sous le signe intégrale)

$$\frac{\partial \mathbb{E}_\theta(\hat{g}_n)}{\partial\theta} = \frac{\partial \mathbb{E}_\mu(\hat{g}_n V_\theta)}{\partial\theta} = \mathbb{E}_\mu\left(\hat{g}_n \frac{\partial V_\theta}{\partial\theta}\right) \quad \forall\theta. \tag{2.10}$$

On dit qu'un tel estimateur est régulier.

Proposition 2.1. *Sous les conditions précédentes*

$$\mathbb{E}_\theta\left(\hat{g}_n - g(\theta)\right)^2 \geq \frac{(g'(\theta))^2}{nI(\theta)} \quad \forall\theta. \tag{2.11}$$

Le membre de droite dans l'inégalité (2.11) s'appelle borne de Cramér-Rao.

Démonstration. L'estimateur est sans biais. Ceci signifie que

$$\mathbb{E}_\theta(\hat{g}_n) \equiv g(\theta) \equiv \mathbb{E}_\mu(\hat{g}_n V_\theta).$$

Donc ces trois fonctions sont dérivables et leur dérivée est $g'(\theta)$. Puisque l'estimateur est régulier, ceci implique, en utilisant (2.5), que

$$g'(\theta) \equiv \mathbb{E}_\mu\left(\hat{g}_n \frac{\partial V_\theta}{\partial\theta}\right) \equiv \mathbb{E}_\theta\left(\hat{g}_n \frac{\partial L_\theta}{\partial\theta}\right) \equiv \mathbb{E}_\theta\left(\left(\hat{g}_n - g(\theta)\right)\frac{\partial L_\theta}{\partial\theta}\right).$$

Puis en utilisant l'inégalité de Schwarz :

$$g'(\theta)^2 \leq \mathbb{E}_\theta\left(\hat{g}_n - g(\theta)\right)^2 \mathbb{E}_\theta\left(\frac{\partial L_\theta}{\partial\theta}\right)^2 = nI(\theta)\mathbb{E}_\theta\left(\hat{g}_n - g(\theta)\right)^2.$$

Ce qui démontre l'inégalité

□

Remarquons le rôle joué par l'information de Fisher : plus grande est cette information, plus la variance du "meilleur" estimateur régulier sans biais est petite.

Définition 2.9.

(1) *On dit qu'un estimateur de $g(\theta)$ est efficace si il est sans biais et si sa variance est égale à la borne de Cramér-Rao.*

(2) *On dit qu'il est asymptotiquement efficace si il est asymptotiqiuement sans biais (ou si il est convergent) et si sa variance est équivalente à cette même borne lorsque n tend vers l'infini.*

Dans l'exemple 2.1, la moyenne arithmétique est un estimateur efficace de θ.
Remarquons aussi que, pour un estimateur régulier, dans un modèle régulier, la variance ne peut tendre vers zéro à une vitesse plus grande que n^{-1}. Il est instructif à cet égard de reprendre l'exemple 2.2 où rien de ce qui précède n'est valable puisque le modèle n'est pas régulier.
Il existe bien sûr une version multivariée de l'inégalité de Cramér-Rao. Pour ne pas alourdir l'exposé, nous l'énoncerons sans démonstration.

PROPOSITION 2.2. *Soit g une application mesurable de $\mathbb{R}^p$ dans $\mathbb{R}^k$. Soit $\hat{g}_n = \hat{g}_n(X_1, \cdots, X_n)$ un estimateur sans biais de $g(\theta) = g(\theta_1, \cdots, \theta_p)$. Supposons que g est différentiable sur Θ. Sous les hypothèses de régularité données dans la définition 2.8, sous la condition (2.10) de régularité de l'estimateur (adaptée bien sûr à la situation multidimensionnelle) la matrice*

$$\mathbb{E}_\theta\left(\left(\hat{g}_n - g(\theta)\right)^t\left(\hat{g}_n - g(\theta)\right)\right) - \begin{pmatrix} \nabla_\theta g_1 \\ \vdots \\ \nabla_\theta g_k \end{pmatrix} I_n^{-1}(\theta) \begin{pmatrix} \nabla_\theta g_1 & \cdots & \nabla_\theta g_k \end{pmatrix}$$

est semi-définie positive.

Ce résultat, appliqué à l'exemple 2.3, en prenant pour g la fonction identité de $\mathbb{R}^2$, indique que pour tout estimateur $(\hat{m}_n, \hat{\sigma}_n^2)$ sans biais et régulier du vecteur (m, σ^2), la matrice

$$\mathbb{E}_{m,\sigma}\left(\begin{pmatrix} \hat{m}_n - m \\ \hat{\sigma}_n^2 - \sigma^2 \end{pmatrix}\begin{pmatrix} \hat{m}_n - m & \hat{\sigma}_n^2 - \sigma^2 \end{pmatrix}\right) - \begin{pmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{2\sigma^4}{n} \end{pmatrix}$$

est semi-définie positive, ce qui donne une limite inférieure pour la variance de n'importe quelle combinaison linéaire des deux estimateurs.

4. Méthode du maximum de vraisemblance

On suppose ici encore que le modèle est dominé.
Une des plus célèbres méthodes d'estimation consiste, $x_1, \cdots, x_n$ ayant été observés, à maximiser (par rapport à θ) la vraisemblance de l'échantillon $V_\theta(x_1, \cdots, x_n)$. L'idée sous-jacente est assez simple. Regardons par exemple le cas où P_θ est une loi discrète. Alors, $V_\theta(x_1, \cdots, x_n) = P_\theta(X_1 = x_1, \cdots, X_n = x_n)$. La procédure consiste donc à choisir, parmi les valeurs du paramètre, celle(s) qui rendent maximum la probabilité d'obtenir précisémment l'échantillon observé, ce qui consiste en somme à dire "si on a observé cet échantillon, c'est qu'il avait une grande probabilité d'apparaître". Dans le cas où P_θ est à densité, la vraisemblance est la densité de l'échantillon, l'interprétation simple qu'on vient de faire ne tient plus puisque toutes les probabilités sont nulles. On peut dire toutefois que, à θ fixé, les observations ont tendance à se grouper dans la (ou les) zones ou la densité est maximale. Ayant observé $(x_1, \cdots, x_n)$, il est naturel de penser que ce vecteur se situe dans une zone où la densité est grande. Ce qui justifie la recherche du paramètre qui maximise la

densité prise au point $(x_1, \cdots, x_n)$. Ces considérations heuristiques étant faites, il se trouve que les estimateurs ainsi construits ont de bonnes propriétés.

4.1. Définition et premières propriétés.

DÉFINITION 2.10. *On appelle estimateur du maximum de vraisemblance de θ toute variable aléatoire $\hat{\theta}_n = \hat{\theta}(X_1, \cdots, X_n)$ à valeurs dans Θ, telle que, $\mu^{\otimes n}$-presque partout,*

$$V_{\hat{\theta}_n}(X_1, \cdots, X_n) \geq V_\theta(X_1, \cdots, X_n), \quad \forall \theta.$$

Trouvez cet estimateur dans l'exemple 2.2.
Notons que, si le modèle est régulier, tout estimateur du maximum de vraisemblance est solution du système des équations de vraisemblance

$$\nabla_\theta L_\theta(X_1, \cdots, X_n) = 0. \tag{2.12}$$

Ecrire et résoudre les équations de vraisemblance dans les exemples 2.1 et 2.3.

4.2. Convergence presque sûre. On va s'intéresser dans un premier temps à la convergence presque sûre. C'est à dire qu'on va montrer que, dans les cas les plus standards, l'estimateur du maximum de vraisemblance converge Π_θ-presque sûrement vers θ lorsque la taille d'échantillon tend vers l'infini.

THÉORÈME 2.3. *Si les variables sont indépendantes, si le modèle est régulier, et si l'application qui à θ associe P_θ est injective, alors il existe une suite $(\hat{\theta}_n)$ de solution des équations de vraisemblance (2.12) qui, quelque soit $\theta \in \Theta$ converge $P_\theta^{\otimes n}$-presque sûrement vers θ lorsque $n \to \infty$.*

DÉMONSTRATION. Pour faciliter les choses on va se restreindre au cas où le paramètre est uni-dimensionnel.
Fixons θ, et considérons $V_t(X_1, \cdots, X_n)$ comme une fonction de t. La maximiser revient à maximiser

$$\frac{V_t(X_1, \cdots, X_n)}{V_\theta(X_1, \cdots, X_n)}.$$

Le maximum de cette fonction, s'il est atteint sur Θ, l'est en un point qui annule la dérivée de

$$\ln \frac{V_t(X_1, \cdots, X_n)}{V_\theta(X_1, \cdots, X_n)} = \sum_{j=1}^{n} \ln \frac{f(X_j, t)}{f(X_j, \theta)}, \tag{2.13}$$

ou, de façon équivalente, qui annule la dérivée de

$$\frac{1}{n} \sum_{j=1}^{n} \ln \frac{f(X_j, t)}{f(X_j, \theta)}.$$

Considérons donc maintenant la suite de variables aléatoires (2.13). Ce sont des variables i.i.d. dont l'espérance est négative (mais éventuellement égale à $-\infty$). En effet, d'après l'inégalité de Jensen :

$$\mathbb{E}_\theta \left(\ln \frac{f(X_j, t)}{f(X_j, \theta)} \right) < \ln \mathbb{E}_\theta \left(\frac{f(X_j, t)}{f(X_j, \theta)} \right) = 0.$$

L'inégalité est stricte : en effet, la fonction ln étant strictement concave on a

$$\mathbb{E}_\theta \left(\ln \frac{f(X, t)}{f(X, \theta)} \right) = \ln \mathbb{E}_\theta \left(\frac{f(X, t)}{f(X, \theta)} \right) \iff \frac{f(X, t)}{f(X, \theta)} = c \quad \mu \, p.s.$$

Comme pour tout t, $f(.,t)$ est une densité, la constante c est nécéssairement égale à 1. Mais $\frac{f(X,t)}{f(X,\theta)} = 1$ μ $p.s.$ contredit l'injectivité de l'application qui à θ associe P_θ. Lorsque

$$\mathbb{E}_\theta \Big| \ln \frac{f(X_j,t)}{f(X_j,\theta)} \Big| < \infty, \tag{2.14}$$

la loi des grands nombres permet de conclure que, P_θ-presque sûrement,

$$\frac{1}{n}\sum_{j=1}^{n} \ln \frac{f(X_j,t)}{f(X_j,\theta)} \to \mathbb{E}_\theta \left(\ln \frac{f(X,t)}{f(X,\theta)} \right) < 0. \tag{2.15}$$

Lorsque (2.14) n'est pas vérifiée alors

$$\mathbb{E}_\theta \left(\left(\ln \frac{f(X_j,t)}{f(X_j,\theta)} \right)^- \right) = \infty, \quad \text{et} \quad \mathbb{E}_\theta \left(\left(\ln \frac{f(X_j,t)}{f(X_j,\theta)} \right)^+ \right) < \infty.$$

Pour vérifier la finitude de la partie positive, on remarque que

$$\begin{aligned}
\mathbb{E}_\theta \left(\left(\ln \frac{f(X_j,t)}{f(X_j,\theta)} \right)^+ \right) &= \int_{f(x,t)>f(x,\theta)} \ln \frac{f(x,t)}{f(x,\theta)} f(x,\theta)\, \mathrm{d}x \\
= \quad & \int_{f(x,t)>f(x,\theta)} |\ln f(x,t)| f(x,\theta)\, \mathrm{d}x + \int_{f(x,t)>f(x,\theta)} |\ln f(x,\theta)| f(x,\theta)\, \mathrm{d}x \\
\leq \quad & \int_{f(x,t)>f(x,\theta)} |\ln f(x,t)| f(x,t)\, \mathrm{d}x + \int_{f(x,t)>f(x,\theta)} |\ln f(x,\theta)| f(x,\theta)\, \mathrm{d}x \\
\leq \quad & \mathbb{E}_t \left(|\ln f(X_j,t)| \right) + \mathbb{E}_\theta \left(|\ln f(X_j,\theta)| \right) < \infty
\end{aligned}$$

d'après (2.3). On montre que P_θ- presque sûrement,

$$\frac{1}{n}\sum_{j=1}^{n} \ln \frac{f(X_j,t)}{f(X_j,\theta)} \to -\infty. \tag{2.16}$$

Considérons maintenant l'ensemble dénombrable

$$\Theta_0 = \{ t \in \Theta | t = \theta \pm \frac{1}{k},\ k \in \mathbb{N}^* \},$$

et, pour tout $t \in \Theta_0$, appelons N_t l'ensemble des $\omega \in \Omega$ tels que sur lequel (2.15) ou (2.16) a lieu, $P_\theta(N_t) = 1$. Évidemment, puisque Θ_0 est dénombrable, l'intersection des N_t est de probabilité P_θ égale à 1. Soit maintenant un ω appartenant à cette intersection. D'après la définition des N_t, pour tout $t \in \Theta_0$, il existe une constante $-\infty \leq l(t) < 0$ telle que,

$$\frac{1}{n}\sum_{j=1}^{n} \ln \frac{f(X_j(\omega),t)}{f(X_j(\omega),\theta)} \to l(t). \tag{2.17}$$

Ayant fixé ε, choisissons $k > \varepsilon^{-1}$ et posons $t_k = \theta - \frac{1}{k}$ et $t'_k = \theta + \frac{1}{k}$. D'après (2.17), il existe un entier $n_0(k,\omega)$ tel que, pour $n \geq n_0(k,\omega)$, on a à la fois

$$\frac{1}{n}\sum_{j=1}^{n} \ln \frac{f(X_j(\omega),t_k)}{f(X_j(\omega),\theta)} < 0 \quad \text{et} \quad \frac{1}{n}\sum_{j=1}^{n} \ln \frac{f(X_j(\omega),t_{k'})}{f(X_j(\omega),\theta)} < 0.$$

Considérons la fonction $\frac{1}{n}\sum_{j=1}^{n} \ln \frac{f(X_j(\omega),t)}{f(X_j(\omega),\theta)}$. Elle est nulle pour $t = \theta$ et strictement négative pour $t = t_k$ et $t = t_{k'}$. Comme elle est partout dérivable, il existe un point de $]t_k, t_{k'}[$ qui annule sa dérivée. On a donc bien prouvé que, pour tout

$n \geq n_0(k,\omega)$, il existe dans $]\theta-\varepsilon,\theta-\varepsilon[$ un point $\hat{\theta}_n(\omega)$ qui est solution des équations de vraisemblance. □

Les hypothèses du théorème ne sont pas du tout nécessaires. Pour s'en persuader il suffit de retourner à l'exemple 2.2 où l'estimateur du maximum de vraisemblance $\max\{X_1,\cdots,X_n\}$ converge P_θ-presque sûrement vers θ alors que le modèle n'est pas régulier.

4.3. Convergence en loi. Nous plaçant dans les conditions du théorème 2.3, nous allons maintenant regarder l'écart renormalisé $\sqrt{n}(\hat{\theta}_n-\theta)$.

THÉORÈME 2.4. *Sous les hypothèses du théorème 2.3, et si de plus*
- *quel que soit x, $f(x,\theta)$ est partout deux fois dérivable par rapport à θ,*
- *les conditions de dérivabilité sous intégrale (2.3) et (2.7) sont vérifiées*
- *la dérivée seconde $\frac{\partial^2 lnf(x,\theta)}{\partial\theta^2}$ est partout continue par rapport à θ, et cette continuité est uniforme en x (on notera $D_2(x,\theta)$ cette fonction)*
- *en tout θ, l'information de Fisher est finie.*

Alors, pour toute suite $(\hat{\theta}_n)$ de solutions des équations de vraisemblance qui converge presque sûrement vers θ, on a sous la loi P_θ,

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

DÉMONSTRATION. Posons

$$h_n(X_1,\cdots,X_n,\theta)=\frac{1}{n}\sum_{j=1}^{n}\frac{\partial lnf(X_j,\theta)}{\partial\theta}.$$

Sous les hypothèses faites, $(\frac{\partial lnf(X_j,\theta)}{\partial\theta})_n$ est une suite i.i.d. de variables aléatoires centrées et de variance $I(\theta)$. On applique le théorème central limite : sous la loi P_θ,

$$\sqrt{n}h_n(X_1,\cdots,X_n,\theta) \to_{\mathcal{L}} \mathcal{N}(0,I(\theta)). \tag{2.18}$$

Puis, développons h_n en utilisant le théorème des accroissements finis :

$$h_n(X_1,\cdots,X_n,t)=h_n(X_1,\cdots,X_n,\theta)+(t-\theta)K_n(X_1,\cdots,X_n,\theta_n^*),$$

où $\theta_n^* \in]t,\theta[$ et

$$K_n(X_1,\cdots,X_n,u)=\frac{1}{n}\sum_{j=1}^{n}D_2(X_j,u).$$

Puis, en remplaçant t par $\hat{\theta}_n$ on a

$$0=h_n(X_1,\cdots,X_n,\theta)+(\hat{\theta}_n-\theta)K_n(X_1,\cdots,X_n,\theta_n^*). \tag{2.19}$$

On sait, en appliquant la loi des grands nombres à la suite i.i.d. $\frac{\partial^2 lnf(X_j,\theta)}{\partial\theta^2}$, que P_θ-presque sûrement,

$$K_n(X_1,\cdots,X_n,\theta)=\frac{1}{n}\sum_{j=1}^{n}D_2(X_j,\theta) \to -I(\theta). \tag{2.20}$$

Par ailleurs,

$$\begin{aligned}|K_n(X_1,\cdots,X_n,\theta_n^*)-K_n(X_1,\cdots,X_n,\theta)| &\leq \frac{1}{n}\sum_{j=1}^{n}|D_2(X_j,\theta_n^*)-D_2(X_j,\theta)| \\ &\leq \sup_x|D_2(x,\theta_n^*)-D_2(x,\theta)|. \end{aligned}\tag{2.21}$$

Comme $\hat{\theta}_n$ converge P_θ-presque sûrement vers θ, il en est de même pour θ_n^*. Grâce à la continuité uniforme en x de la dérivée seconde $D_2(x,\theta)$, il résulte de (2.21) que, P_θ-presque sûrement,

$$K_n(X_1,\cdots,X_n,\theta_n^*) \to -I(\theta).$$

Utilisant ce dernier résultat et (2.19), le théorème est prouvé. □

Retourner à l'exemple 2.2 et trouver la loi limite de l'écart $(\hat{\theta}_n - \theta)$ après l'avoir convenablement renormalisé.

5. Exhaustivité

On appelle statistique n'importe quelle application mesurable définie sur l'échantillon. Ceci sous-entend que, comme c'était le cas pour les estimateurs, cette statistique ne dépend pas du paramètre inconnu θ.

5.1. Définition et critère. Grosso-modo, une statistique exhaustive contient toute l'information sur le paramètre apportée par l'échantillon. On devrait donc retrouver (vérifiez-le) dans les définitions équivalentes qui suivent, que l'échantillon lui-même est une statistique exhaustive. Bien entendu elle est inintéressante, et les situations agréables sont celles ou l'on trouve une statistique qui est un "vrai résumé" exhaustif (voir les exemples).

DÉFINITION 2.11. *Une application mesurable S de $\underline{X}^n$ dans $\mathbb{R}^m$ est exhaustive si pour toute $A \in \mathcal{X}^{\otimes n}$, il existe une version de*

$$P_\theta\left((X_1,\cdots,X_n) \in A | S\right)$$

qui ne dépend pas de θ, ce qui revient à dire que pour toute application g mesurable et $P_\theta^{\otimes n}$-intégrable de $\underline{X}^n$ dans $\mathbb{R}$, il existe une version de

$$\mathbb{E}_\theta\left(g(X_1,\cdots,X_n)|S\right)$$

qui ne dépend pas de θ.

Cette définition peut être interprétée de la façon suivante : une fois connue la valeur de S, la loi de l'échantillon ne dépend plus de θ. De cette façon on comprend que S porte toute l'information sur le paramètre. Ce qui suit donne un critère très commode pour vérifier qu'une statistique est exhaustive par simple inspection de la vraisemblance.

PROPOSITION 2.5. *Supposons le modèle dominé. La statistique S, à valeurs dans $\mathbb{R}^m$, est exhaustive si et seulement si il existe h_θ et ϕ, deux fonctions positives respectivement définies sur $\mathbb{R}^m$ et sur $\underline{X}^n$ telles que $\mu^{\otimes n}$-presque partout,*

$$V_\theta(x_1,\cdots,x_n) = \phi(x_1,\cdots,x_n) h_\theta\left(S(x_1,\cdots,x_n)\right). \tag{2.22}$$

DÉMONSTRATION. On montrera seulement que la condition est suffisante : si (2.22) est vérifiée, S est exhaustive.
Soit g une aplication mesurable et P_θ-intégrable de $\underline{X}^n$ dans $\mathbb{R}$ et $\mathcal{B}$ une sous tribu de $\mathcal{X}^{\otimes n}$. On a P_θ-p.s.

$$\mathbb{E}_\theta(g(X_1,\cdots,X_n)|\mathcal{B}) = \frac{\mathbb{E}_\mu(V_\theta(X_1,\cdots,X_n)g(X_1,\cdots,X_n)|\mathcal{B})}{\mathbb{E}_\mu(V_\theta(X_1,\cdots,X_n)|\mathcal{B})}\mathbb{I}_{\{\mathbb{E}_\mu(V_\theta|\mathcal{B})\neq 0\}}. \tag{2.23}$$

Pour prouver (2.23), posons $C = \{\mathbb{E}_\mu(V_\theta|\mathcal{B}) \neq 0\}$. On remarque que C est $\mathcal{B}$-mesurable et

$$P_\theta(C^c) = \mathbb{E}_\mu(V_\theta \mathbb{I}_{C^c}) = \mathbb{E}_\mu\left(\mathbb{I}_{C^c}\mathbb{E}_\mu(V_\theta|\mathcal{B})\right) = 0.$$

Ensuite, si $B \in \mathcal{B}$,

$$\begin{aligned}\int_B \mathbb{I}_C \frac{\mathbb{E}_\mu(V_\theta g|\mathcal{B})}{\mathbb{E}_\mu(V_\theta|\mathcal{B})} dP_\theta &= \int_{B\cap C} \frac{\mathbb{E}_\mu(V_\theta g|\mathcal{B})}{\mathbb{E}_\mu(V_\theta|\mathcal{B})} V_\theta d\mu = \mathbb{E}_\mu\left(\mathbb{I}_{B\cap C}\frac{\mathbb{E}_\mu(V_\theta g|\mathcal{B})}{\mathbb{E}_\mu(V_\theta|\mathcal{B})}V_\theta\right)\\ &= \mathbb{E}_\mu\left(\mathbb{I}_{B\cap C}\frac{\mathbb{E}_\mu(V_\theta g|\mathcal{B})}{\mathbb{E}_\mu(V_\theta|\mathcal{B})}\mathbb{E}_\mu(V_\theta|\mathcal{B})\right) = \mathbb{E}_\mu\left(\mathbb{I}_{B\cap C}\mathbb{E}_\mu(V_\theta g|\mathcal{B})\right)\\ &= \mathbb{E}_\mu(\mathbb{I}_{B\cap C}V_\theta g) = E_\theta(\mathbb{I}_{B\cap C}g) = E_\theta(\mathbb{I}_B g),\end{aligned}$$

ce qui prouve (2.23). On applique ce résultat à la tribu engendrée par S, et on utilise l'expression (2.22) :

$$\begin{aligned}\mathbb{E}_\theta(g|S) &= \frac{\mathbb{E}_\mu(\phi h_\theta(S) g|S)}{\mathbb{E}_\mu(\phi h_\theta(S))|S)}\mathbb{I}_{\{\mathbb{E}_\mu(\phi h_\theta(S)|S)\neq 0\}}\\ &= \frac{h_\theta(S)\mathbb{E}_\mu(\phi g|S)}{h_\theta(S)\mathbb{E}_\mu(\phi|S)}\mathbb{I}_{\{h_\theta(S)\neq 0\}}\mathbb{I}_{\{\mathbb{E}_\mu(\phi|S)\neq 0\}}\\ &= \frac{\mathbb{E}_\mu(\phi g|S)}{\mathbb{E}_\mu(\phi|S)}\mathbb{I}_{\{\mathbb{E}_\mu(\phi|S)\neq 0\}} \quad P_\theta - p.s.\end{aligned}$$

ce qui termine la preuve. □

5.2. Statistique exhaustive et estimateur du maximum de vraisemblance. Supposons qu'il existe une décomposition de type (2.22) de la vraisemblance. Alors, évidemment, maximiser la vraisemblance revient à maximiser $h_t(S(X_1, \cdots, X_n))$. Or le maximum de cette fonction de t, si il est atteint, l'est en un point $\hat{\theta}_n$ qui est fonction de $S(X_1, \cdots, X_n)$. D'où

PROPOSITION 2.6. *Si S, à valeurs dans $\mathbb{R}^m$, est exhaustive et si $\hat{\theta}_n(X_1, \cdots, X_n)$ est un estimateur du maximum de vraisemblance de θ, il existe une application mesurable g_n de $\mathbb{R}^m$ dans Θ telle que*

$$\hat{\theta}_n(X_1, \cdots, X_n) = g_n(S(X_1, \cdots, X_n))$$

5.3. Exhaustivité et information. Nous supposons ici que le modèle est régulier et que l'information de Fisher est finie. Tout comme on a défini l'information apportée par une observation et l'information apportée par l'échantillon, on peut définir l'information apportée par une statistique $S(X_1, \cdots, X_n)$ quelconque.
Soit Q_θ la loi de S, c'est à dire la transportée par S de la mesure produit $P_\theta^{\otimes n}$. Soit aussi ν la transportée par S de la mesure dominante μ. Ce sont des mesures sur $\mathbb{R}^m$. Il est facile de voir que les Q_θ sont équivalentes à ν : en effet

$$Q_\theta(A) = 0 \iff P_\theta^{\otimes n}(S^{-1}(A)) = 0 \iff \mu(S^{-1}(A)) = 0 \iff \nu(A) = 0.$$

Donc il y a une densité de Q_θ par rapport à ν. Cette densité est strictement positive. Nous la noterons $q(s, \theta)$.

DÉFINITION 2.12. *On appelle information apportée par la statistique S la quantité (éventuellement infinie)*

$$I_\theta^S = \mathbb{E}_\theta\left(\frac{\partial \ln q(S,\theta)}{\partial\theta}\right)^2$$

Notons que, d'après la remarque 7 de la section 2.2, l'information I_θ^S ne change pas si la mesure ν est changée en une mesure $\nu^\star$ équivalente.
Si on suppose que le modèle $(\mathbb{R}^m, \mathcal{B}(\mathbb{R}^m), Q_\theta)$ est régulier, l'information apportée par une statistique ne peut dépasser celle apportée par l'échantillon.

PROPOSITION 2.7. *Si le modèle image par la statistique S et le modèle de départ sont tous deux réguliers, et si les informations I_θ^S et $I_n(\theta)$ sont finies,*

$$I_\theta^S \leq I_n(\theta). \tag{2.24}$$

L'égalité a lieu lorsque S est exhaustive.

DÉMONSTRATION. Montrons d'abord que

$$\frac{\partial \ln q(S,\theta)}{\partial \theta} = \mathbb{E}_\theta \left(\frac{\partial L_\theta(X_1, \cdots, X_n)}{\partial \theta} \Big| S \right). \tag{2.25}$$

Pour cela, pour tout $A \in \mathcal{B}(\mathbb{R}^m)$,

$$\begin{aligned}
\int_A \frac{\partial \ln q(s,\theta)}{\partial \theta} dQ_\theta(s) &= \int_A \frac{\partial q(s,\theta)}{\partial \theta} d\nu(s) \\
&= \frac{\partial}{\partial \theta} \int_A q(s,\theta) d\nu(s) = \frac{\partial P_\theta(S \in A)}{\partial \theta} \\
&= \frac{\partial}{\partial \theta} \int_{S^{-1}(A)} V_\theta d\mu = \int_{S^{-1}(A)} \frac{\partial V_\theta}{\partial \theta} d\mu \\
&= \int_{S^{-1}(A)} \frac{\partial L_\theta}{\partial \theta} dP_\theta.
\end{aligned}$$

donc (2.25) est prouvée.
Mais alors

$$\mathrm{Var}_\theta \left(\frac{\partial \ln q(S,\theta)}{\partial \theta} \right) \leq \mathrm{Var}_\theta \left(\frac{\partial L_\theta(X_1, \cdots, X_n)}{\partial \theta} \right),$$

ce qui est l'inégalité (2.24).
Supposons maintenant que S est exhaustive. Reprenons la décomposition (2.22) de la vraisemblance. Vu les hypothèses, la fonction ϕ est ($\mu^{\otimes n}$-presque sûrement) partout strictement positive. Soit $\mu^\star$ la mesure dont la densité est ϕ par rapport à $\mu^{\otimes n}$. Soit $\nu^\star$ son image par S. Il est facile de voir que $h_\theta(s)$ est la densité de Q_θ, loi de S par rapport à $\nu^\star$: en effet,

$$\begin{aligned}
\int_A h_\theta(s) d\nu^\star(s) &= \int_{S^{-1}(A)} h_\theta(S(x_1, \cdots, x_n)) d\mu^\star(x_1, \cdots, x_n) \\
&= \int_{S^{-1}(A)} h_\theta(S(x_1, \cdots, x_n)) \phi(x_1, \cdots, x_n) d\mu(x_1, \cdots, x_n) \\
&= \int_{S^{-1}(A)} dP_\theta^{\otimes n}(x_1, \cdots, x_n) = P_\theta(S \in A) = Q_\theta(A).
\end{aligned}$$

Par définition, on a

$$\begin{aligned} I^S(\theta) &= \mathbb{E}_\theta \left(\frac{\partial \ln h_\theta(S)}{\partial \theta} \right)^2 \\ &= \mathbb{E}_\theta \left(\frac{\partial \ln \Big(h_\theta(S(X_1, \cdots, X_n)) \phi(X_1, \cdots, X_n) \Big)}{\partial \theta} \right)^2 \\ &= \mathbb{E}_\theta \left(\frac{\partial L_\theta(X_1, \cdots, X_n)}{\partial \theta} \right)^2 = I_n(\theta). \end{aligned}$$

La proposition est prouvée. □

5.4. Amélioration d'un estimateur sans biais grâce à une statistique exhaustive. En fait, les meilleurs estimateurs sans biais sont à rechercher parmi les fonctions de statistiques exhaustives.

THÉORÈME 2.8. *Théorème de Rao-Blackwell*
Soit $\hat{g}_n$ un estimateur sans biais de $g(\theta)$ et soit S une statistique exhaustive. Alors, $\mathbb{E}_\theta(\hat{g}_n|S) = \mathbb{E}(\hat{g}_n|S)$ est un estimateur sans biais de $g(\theta)$, et $\mathbb{E}(\hat{g}_n|S) \gg \hat{g}_n$. Autrement dit, pour tout $\theta \in \Theta$

$$\begin{aligned} \mathbb{E}_\theta \left(\mathbb{E}(\hat{g}_n|S) \right) &= g(\theta) \\ \mathrm{Var}_\theta \left(\mathbb{E}(\hat{g}_n|S) \right) &\leq \mathrm{Var}_\theta(\hat{g}_n). \end{aligned}$$

DÉMONSTRATION. Il est important de remarquer d'abord que $\mathbb{E}_\theta(\hat{g}_n|S)$ peut bien être considéré comme une estimateur. En effet, $\mathbb{E}_\theta(\hat{g}_n|S)$ ne dépend pas de θ parce que S est exhaustive. C'est d'ailleurs pourquoi l'indice θ est omis dans son écriture. Le reste du théorème est conséquence directe des propriétés élémentaires de l'espérance conditionnelle. □

Il faut remarquer que, bien que l'estimateur ainsi construit ait une variance inférieure ou égale à celle de celui dont on est parti, rien ne garantit qu'il soit admissible. Notamment rien ne garantit qu'il atteigne la borne de Cramér-Rao.

5.5. Statistique totale (ou complète).

DÉFINITION 2.13. *Soit une statistique S à valeurs dans $\mathbb{R}^m$, de loi Q_θ. On dit qu'elle est totale si, pour toute fonction ϕ mesurable définie sur $\mathbb{R}^m$ telle que $\mathbb{E}_\theta|\phi(S)| < \infty$,*

$$\{\mathbb{E}_\theta \phi(S) = 0 \quad \forall \theta \in \Theta\} \Longrightarrow \{\phi = 0, \quad Q_\theta - p.s. \quad \forall \theta \in \Theta\}.$$

Il n'est pas toujours facile de vérifier qu'une statistique est totale. Par exemple la somme $\sum_{j=1}^n X_j$ est totale pour le modèle de Poisson et le max pour le modèle Uniforme sur $[0, \theta]$.

THÉORÈME 2.9. *Si $\hat{g}_n$ est un estimateur sans biais de $g(\theta)$ et si S est exhaustive et totale, alors, l'amélioré de Rao-Blackwell $\mathbb{E}_\theta(\hat{g}_n|S) = \mathbb{E}(\hat{g}_n|S)$ est optimal dans la classe des estimateurs sans biais de $g(\theta)$.*

DÉMONSTRATION. l'énoncé signifie que, si Z est un estimateur sans biais de $g(\theta)$, on a $\mathbb{E}(\hat{g}_n|S) \gg Z$, c'est à dire

$$\mathrm{Var}_\theta Z \geq \mathrm{Var}_\theta \left(\mathbb{E}(\hat{g}_n|S) \right) \quad \forall \theta \in \Theta.$$

Prouvons le. On a d'après le théorème de Rao-Blackwell,

$$\mathrm{Var}_\theta Z \geq \mathrm{Var}_\theta \left(\mathbb{E}(Z|S)\right) \quad \forall \theta \in \Theta.$$

Mais on a aussi pour tout θ, $\mathbb{E}_\theta \left(\mathbb{E}(Z|S) - \mathbb{E}(\hat{g}_n|S)\right) = 0$. Autrement dit, en prenant $\phi(s) = \mathbb{E}(Z|S=s) - \mathbb{E}(\hat{g}_n|S=s)$,

$$\mathbb{E}_\theta \phi(S) = 0 \quad \forall \theta \in \Theta.$$

Comme S est totale, on en déduit que

$$\phi = 0, \quad Q_\theta - p.s. \quad \forall \theta \in \Theta,$$

C'est à dire que les deux estimateurs $\mathbb{E}(Z|S)$ et $\mathbb{E}(\hat{g}_n|S)$ sont Q_θ-p.s. les mêmes, ce qui termine la preuve. □

CHAPITRE 3

Tests d'hypothèses

1. Idées générales, définitions et un exemple

Le cadre est le même que dans l'introduction. On suppose en outre qu'on dispose de deux parties disjointes Θ_0 et Θ_1 de Θ. On souhaite en se basant sur le n-échantillon $(X_1, \cdots, X_n)$, décider si la valeur (inconnue) du paramètre θ qui gouverne la loi du n-échantillon est dans Θ_0 ou plutôt dans Θ_1.

(1) L'hypothèse $\{\theta \in \Theta_0\}$, notée encore parfois H_0, est appelée *hypothèse nulle*

(2) L'hypothèse $\{\theta \in \Theta_1\}$, notée encore parfois H_1, est appelée *hypothèse alternative* ou aussi *contre-hypothèse*.

(3) On dit qu'on teste H_0 contre H_1.

(4) Si l'un des ensembles précédents est réduit à un point on dit que l'hypothèse correspondante est simple.

Quelle que soit la procédure choisie, elle pourra mener à deux types d'erreurs :

Type I: rejeter l'hypothèse nulle alors qu'en fait $\theta \in \Theta_0$: cette erreur s'appelle *erreur de première espèce*,

Type II: accepter l'hypothèse nulle alors qu'en fait $\theta \in \Theta_1$: c'est l'*erreur de deuxième espèce*.

En première analyse, un test ne sera rien d'autre que le choix d'une partition de $\underline{X}^n$ en deux régions complémentaires

R_0**:** *région de rejet* ou *région critique*

R_0^c**:** *région d'acceptation*.

Si $(X_1, \cdots, X_n) \in R_0$, le test rejette H_0, tandis qu'il accepte H_0 si $(X_1, \cdots, X_n) \in R_0^c$.

On peut donc définir deux probabilités d'erreur :

Type I: Lorsque $\theta \in \Theta_0$, $P_\theta\left((X_1, \cdots, X_n) \in R_0\right)$ représente la *probabilité d'erreur de première espèce*

Type II: Lorsque $\theta \in \Theta_1$, $P_\theta\left((X_1, \cdots, X_n) \in R_0^c\right)$ représente la *probabilité d'erreur de deuxième espèce*.

DÉFINITION 3.1. *(Niveau, puissance et biais.)*

1) On appelle niveau du test la quantité

$$\alpha = \sup\{P_\theta\left((X_1, \cdots, X_n) \in R_0\right) | \theta \in \Theta_0\}$$

2) On appelle puissance du test l'application définie sur Θ_1

$$P_\theta\left((X_1, \cdots, X_n) \in R_0\right) \quad \theta \in \Theta_1,$$

3) On dit que le test est sans biais si

$$P_\theta\left((X_1,\cdots,X_n)\in R_0\right)\geq\alpha\quad\forall\theta\in\Theta_1,$$

L'idée qui préside à la construction de tests est la plupart du temps : fixer le niveau à une valeur (petite !) α et trouver un test de niveau α qui ait une puissance assez grande. Évidemment une idée intéressante est d'utiliser, pour construire ce test, un bon estimateur $\hat{\theta}_n$ de θ si on en connaît un. Dans ce cas la région critique portera sur $\hat{\theta}_n$.

EXEMPLE 3.1. On jette une pièce 5 fois et elle tombe 5 fois sur pile. Que conclure ? Sous-entendu, est elle truquée ? Même question si elle tombe 4 fois sur pile.
Soit p la probabilité que la pièce tombe sur pile : l'hypothèse H_0 dans ce problème peut être considérée comme $\{p=1/2\}$ et l'hypothèse H_1 comme $\{p>1/2\}$ (la pièce est truquée).
Il est raisonnable, appelant N la variable aléatoire égale au nombre de pile obtenus, de baser le test sur N. Cela veut dire que l'on décide a priori que le test aura la forme suivante : si $N\geq c$ on décide que $p>1/2$, et si $N<c$ on décide que $p=1/2$. Autrement dit la région de rejet est $\{N\geq c\}=R_0$.
La question qui reste à résoudre est de déterminer c. Accessoirement on pourra revenir aux valeurs qui ont servi de prétexte de départ ($N=5$ et $N=4$) et conclure le test dans ces deux cas.
Pour déterminer c on fixe un niveau α et on cherche c tel que

$$P_{p=1/2}(N\geq c)=\alpha.$$

Ici on se heurte clairement à une difficulté illustrée dans le tableau ci-dessous. La variable N étant à valeurs entières, le niveau $P_{p=1/2}(N\geq c)$ ne varie pas lorsque c varie entre deux entiers successifs. Prenons $c=4$ puis $c=5$.

	$c=4$	$c=5$
Niveau	$P_{p=1/2}(N\geq 4)=6/32$	$P_{p=1/2}(N\geq 5)=1/32$
Puissance	$P_p(N\geq 4)=p^5+5p^4(1-p)$	$P_p(N\geq 5)=p^5$

On voit d'abord que la puissance, dans les deux cas, est croissante. De ce fait, on a dans les deux cas $P_p(N\geq c)\geq P_{1/2}(N\geq c)=\alpha$. En d'autres termes, les deux tests sont sans biais.
On remarque aussi que le niveau passe sans transition de $6/32$ pour le premier test à $1/32$ pour le second. Notamment le niveau 5% ne sera jamais atteint.
Ceci étant, si on observe 5 pile, pour un test de niveau $1/32$ aussi bien que pour un test de niveau $6/32$ on conclut que la pièce est truquée. ‖

2. Test randomisé et Lemme de Neyman-Pearson

Pour résoudre le problème soulevé par l'exemple qui précède, il suffit de modifier légèrement la définition d'un test.

2.1. Test randomisé. Au lieu de dire qu'une procédure de test consiste à rejeter l'hypothèse H_0 dans une certaine région R_0 et l'accepter dans la région complémentaire, on va convenir qu'effectuer un test consiste, en chaque point de $(X_1,\cdots,X_n)\in\underline{X}^n$, à rejeter H_0 *avec une certaine probabilité* $\Phi(X_1,\cdots,X_n)$ et à l'accepter avec la probabilité $1-\Phi(X_1,\cdots,X_n)$.

Un test est, dans ce nouveau point de vue, une application Φ de $\underline{X}^n$ dans $[0,1]$. Φ s'appelle la fonction de test.

La définition donnée dans la section 1 correspond au cas particulier où la fonction de test Φ est égale à $\mathbb{I}_{R_0}$, elle ne prend que les valeurs 0 et 1.

on dit que le test est randomisé lorsque $\Phi(\underline{X}^n) \neq \{0,1\}$.

Bien sûr, il convient de re-définir les notions de niveau et de puissance de façon adéquate.

DÉFINITION 3.2. *(Niveau, puissance et biais.)*

(1) On appelle niveau du test randomisé la quantité

$$\alpha = \sup\{\mathbb{E}_\theta \Phi(X_1, \cdots, X_n) | \theta \in \Theta_0\}$$

(2) On appelle puissance du test randomisé l'application définie sur Θ_1

$$\mathbb{E}_\theta \Phi(X_1, \cdots, X_n) \quad \theta \in \Theta_1,$$

(3) On dit que le test est sans biais si

$$\mathbb{E}_\theta \Phi(X_1, \cdots, X_n) \geq \alpha \quad \forall \theta \in \Theta_1,$$

EXEMPLE 3.2. Reprenons l'exemple 3.1. Supposons qu'on désire construire un test de niveau $\alpha = 5\%$. Il suffit de rejeter H_0 avec une probabilité 1 lorsque $N = 5$ et avec une probabilité γ lorsque $N = 4$, ce qui revient à prendre

$$\Phi(N) = \mathbb{I}_{\{5\}}(N) + \gamma \mathbb{I}_{\{4\}}(N).$$

On ajuste la valeur de γ en écrivant que le niveau est 5%, soit

$$\frac{1}{32} + \frac{5\gamma}{32} = \frac{5}{100},$$

soit $\gamma = 3/25$.

Il est facile de vérifier que le test est sans biais. ‖

2.2. Lemme de Neymann-Pearson. On va ici considérer le cas très particulier où les deux hypothèses sont simples. On teste $\theta = \theta_0$ contre $\theta = \theta_1$.

Ce modèle est toujours dominé par $\mu = P_{\theta_0} + P_{\theta_1}$. On note $V_{\theta_0}(X_1, \cdots, X_n)$ et $V_{\theta_1}(X_1, \cdots, X_n)$ les vraisemblances respectives sous la loi P_{θ_0} et sous la loi P_{θ_1}.

Nous allons considérer tout particulièrement les fonctions de test de la forme

$$\Phi = \mathbb{I}_{\{V_{\theta_1} > cV_{\theta_0}\}} + \gamma \mathbb{I}_{\{V_{\theta_1} = cV_{\theta_0}\}}. \tag{3.1}$$

Ces tests s'appellent *tests du rapport de vraisemblance*, car ils consistent grossomodo à choisir θ_0 ou θ_1 selon que le rapport $V_{\theta_1} V_{\theta_0}^{-1}$ est petit ou grand. Comme dans l'exemple précédent, le fait de choisir a priori un test randomisé vient du souci d'obtenir exactement le niveau souhaité.

THÉORÈME 3.1. *Lemme de Neyman-Pearson. Soit* $\alpha \in]0,1[$. *Il existe* $\gamma \in [0,1]$ *et* $c \geq 0$ *tels que le test du rapport de vraisemblance (3.1) correspondant a les propriétés suivantes :*

1– $\mathbb{E}_{\theta_0} \Phi(X_1, \cdots, X_n) = \alpha$,

2– $\mathbb{E}_{\theta_1} \Phi(X_1, \cdots, X_n) \geq \alpha$,

3– Pour toute fonction de test Φ' *telle que* $\mathbb{E}_{\theta_0} \Phi'(X_1, \cdots, X_n) \leq \alpha$ *on a*

$$\mathbb{E}_{\theta_1} \Phi'(X_1, \cdots, X_n) \leq \mathbb{E}_{\theta_1} \Phi(X_1, \cdots, X_n).$$

Autrement dit il existe un test du rapport de vraisemblance de niveau α, ce test est sans biais et il est le plus puissant parmi les tests de niveau au plus α pour tester $\theta = \theta_0$ contre $\theta = \theta_1$.

DÉMONSTRATION.

Propriété 1.

On remarque d'abord que $P_{\theta_0}(V_{\theta_0} = 0) = 0$. Notons alors $C = \{V_{\theta_0} \neq 0\}$. On a

$$\Psi(z) = P_{\theta_0}(V_{\theta_1} > zV_{\theta_0}) = P_{\theta_0}(\frac{V_{\theta_1}}{V_{\theta_0}}\mathbb{I}_C > z) = 1 - P_{\theta_0}(\frac{V_{\theta_1}}{V_{\theta_0}}\mathbb{I}_C \leq z).$$

Comme $P_{\theta_0}(\frac{V_{\theta_1}}{V_{\theta_0}}\mathbb{I}_C \leq z)$ est une fonction de répartition, on sait qu'en tout z elle est continue à droite et admet une limite à gauche. Il en est donc de même pour la fonction Ψ. Cette fonction est décroissante et on a :

1) $\Psi(z) = 1$ si $z < 0$,
2) $\Psi(0) = P_{\theta_0}(V_{\theta_1} > 0)$,
3) $\Psi(z) \to 0$ quand $z \to +\infty$.

Soit maintenant $c = \inf\{z \geq 0 | \Psi(z) < \alpha\}$. On a

$$\Psi(c) \leq \alpha \leq \Psi(c^-). \tag{3.2}$$

Deux cas peuvent se produire :

situation 1 : Ψ est continue au point c. Alors il y a égalité dans (3.2), et le test défini par $\Phi = \mathbb{I}_{\{V_{\theta_1} > cV_{\theta_0}\}}$, c'est à dire le test non randomisé qui a pour région critique $\{V_{\theta_1} > cV_{\theta_0}\}$ est de niveau exactement α.

situation 2: la fonction Ψ a un saut au point c. Ce saut est d'amplitude $\Psi(c^-) - \Psi(c) = P_{\theta_0}(V_{\theta_1} = cV_{\theta_0})$. Choisissons

$$\gamma = \frac{\alpha - \Psi(c)}{\Psi(c^-) - \Psi(c)}.$$

On en déduit

$$\alpha = \Psi(c) + \gamma(\Psi(c^-) - \Psi(c)) = P_{\theta_0}(V_{\theta_1} > cV_{\theta_0}) + \gamma P_{\theta_0}(V_{\theta_1} = cV_{\theta_0}),$$

et de nouveau on obtient un test de niveau α.

Propriété 3

Soit un test Φ' de niveau au plus α. On a

$$\begin{aligned}\mathbb{E}_{\theta_1}(\Phi' - \Phi) &= \int_{\underline{X}^n} (\Phi' - \Phi) V_{\theta_1} d\mu^{\otimes n} \\ &= \int_{A_1} (\Phi' - \Phi) V_{\theta_1} d\mu^{\otimes n} + \int_{A_2} (\Phi' - \Phi) V_{\theta_1} d\mu^{\otimes n} + \int_{A_3} (\Phi' - \Phi) V_{\theta_1} d\mu^{\otimes n},\end{aligned}$$

où

$$A_1 = \{V_{\theta_1} > cV_{\theta_0}\}, \quad A_2 = \{V_{\theta_1} = cV_{\theta_0}\}, \text{ et } A_3 = \{V_{\theta_1} < cV_{\theta_0}\}.$$

Sur A_1 on a $\Phi = 1$ donc $\Phi' - \Phi \leq 0$ et

$$\int_{A_1} (\Phi' - \Phi) V_{\theta_1} d\mu^{\otimes n} \leq c \int_{A_1} (\Phi' - \Phi) V_{\theta_0} d\mu^{\otimes n}$$

Sur A_3 on a $\Phi = 0$ donc $\Phi' - \Phi \geq 0$ et

$$\int_{A_3} (\Phi' - \Phi) V_{\theta_1} d\mu^{\otimes n} \leq c \int_{A_3} (\Phi' - \Phi) V_{\theta_0} d\mu^{\otimes n}.$$

De plus,

$$\int_{A_2} (\Phi' - \Phi) V_{\theta_1} d\mu^{\otimes n} = c \int_{A_2} (\Phi' - \Phi) V_{\theta_0} d\mu^{\otimes n}.$$

En résumé,

$$\mathbb{E}_{\theta_1}(\Phi' - \Phi) \leq c\mathbb{E}_{\theta_0}(\Phi' - \Phi) = c\left(\mathbb{E}_{\theta_0}(\Phi') - \alpha\right) \leq 0.$$

Propriété 2. Elle se déduit facilement de la propriété 3. Pour cela, considérons la fonction de test $\Phi'(X_1, \cdots, X_n) \equiv \alpha$. D'après ce qu'on vient de voir, ce test qui est de niveau α a une puissance inférieure ou égale à celle de Φ. Mais la puissance de Φ' est égale à α ce qui termine la preuve. □

2.3. Tests du rapport de vraisemblance et exhaustivité. Une remarque s'impose :
soit S une statistique exhaustive. Alors, utilisant la décomposition (2.22) de la vraisemblance on conclut que

$$\mathbb{I}_{V_{\theta_1} > cV_{\theta_0}} + \gamma \mathbb{I}_{V_{\theta_1} = cV_{\theta_0}} \iff \mathbb{I}_{h_{\theta_1}(S) > ch_{\theta_0}(S)} + \gamma \mathbb{I}_{h_{\theta_1}(S) = ch_{\theta_0}(S)}$$

si bien que la fonction de test Φ ne dépend que de S. Autrement dit, *Les tests du rapport de vraisemblance sont basés sur la statistique exhaustive.*
En particulier, dans l'exemple traité plus haut, le nombre N de pile est une statistique exhaustive. On vérifiera que le test randomisé que nous avons utilisé est précisément le test de Neyman-Pearson.

3. Tests sur les espérances d'un échantillon gaussien

Revenons aux exemples 1.1 et 1.2 de l'introduction de ce cours : le taux de cholestérol mesuré sur 200 femmes âgées de 50 ans et le taux d'une certaine hormone mesuré sur 40 personnes avant traitement médical puis après traitement médical. On désire, à l'aide d'un test statistique, tirer une conclusion sur le taux de cholestérol (par exemple est il supérieur à 3.4) ou sur l'efficacité du traitement. Pour cela on va supposer que les variables sous-jacentes (voir l'introduction) sont gaussiennes. Dans ce qui suit ϕ désigne la densité de la loi gaussienne standard et pour tout $\alpha \in]0, 1[$, z_α est défini par

$$\int_{z_\alpha}^{\infty} \phi(x) dx = \alpha.$$

3.1. Tests de comparaison à une valeur donnée de l'espérance d'un échantillon gaussien de variance connue. Soit $(X_1, \cdots, X_n)$ un n-échantillon de la loi $\mathcal{N}(m, \sigma^2)$. On va effectuer successivement, le paramètre σ^2 étant connu, un test de deux hypothèses simples

$$H_0 : \ m = m_0 \quad \text{contre} \quad H_1 : \ m = m_1,$$

un test unilatéral

$$H_0 : \ m \leq m_0 \quad \text{contre} \quad H_1 : \ m > m_0,$$

un test bilatéral

$$H_0 : \ m = m_0 \quad \text{contre} \quad H_1 : \ m \neq m_0.$$

3.1.1. *Test de deux hypothèses simples.* .

Pour fixer les idées, on suppose $m_0 < m_1$. Le rapport des vraisemblances est

$$\begin{aligned}\frac{f_{m_1}(x_1,\cdots,x_n)}{f_{m_0}(x_1,\cdots,x_n)} &= \exp -\frac{\sum_{j=1}^n (X_j-m_1)^2 - \sum_{j=1}^n (X_j-m_0)^2}{2\sigma^2} \\ &= \exp -\frac{n(m_1^2-m_0^2)+2\sum_{j=1}^n X_j(m_0-m_1)}{2\sigma^2}.\end{aligned}$$

Comme $m_0 - m_1 < 0$, la région critique du test de Neyman-Pearson a la forme

$$\sum_{j=1}^n X_j \geq c.$$

(On aura préalablement remarqué que la randomisation n'a pas lieu d'être car $\sum_{j=1}^n X_j$ est gaussienne, ce qui implique que $P(\sum_{j=1}^n X_j = c) = 0$).
On ajuste c pour obtenir un niveau fixé $\alpha \in]0,1[$. Pour cela, on écrit que

$$\alpha = P_{m_0}(\sum_{j=1}^n X_j \geq c) = P\left(\frac{\sum_{j=1}^n X_j - nm_0}{\sqrt{n\sigma^2}} \geq \frac{c-nm_0}{\sqrt{n\sigma^2}}\right),$$

d'où

$$c = nm_0 + z_\alpha \sigma\sqrt{n}.$$

La région critique du test de Neyman-Pearson de niveau α est donc

$$\overline{X}_n \geq m_0 + z_\alpha \frac{\sigma}{\sqrt{n}}. \tag{3.3}$$

La puissance de ce test est

$$\begin{aligned}P_{m_1}\left(\overline{X}_n \geq m_0 + z_\alpha\frac{\sigma}{\sqrt{n}}\right) &= P\left(\sqrt{n}\frac{\overline{X}_n - m_1}{\sigma} \geq \sqrt{n}\frac{m_0-m_1}{\sigma} + z_\alpha\right) \\ &= \int_{z_\alpha+\sqrt{n}\frac{m_0-m_1}{\sigma}}^{\infty} \phi(x)dx,\end{aligned}$$

qui tend vers 1 lorsque n tend vers l'infini.

Remarque 3.1. On remarque que, quelle que soit la contre hypothèse $m = m_1 > m_0$, la région critique du test de Neyman-Pearson reste (3.3). Ceci conduit à utiliser cette région critique pour d'autres tests.
|

Exercice 1. Construire le test du rapport de vraisemblance lorsque $m_1 < m_0$.
△△

3.1.2. *Test unilatéral.*

Définition 3.3. *test UMP "uniformly most powerful"*

Soit un test de niveau α pour tester Θ_0 contre Θ_1. Soit Φ sa fonction de test. On dit que ce test est uniformément le plus puissant (UMP) parmi les tests de niveau au plus α si, pour tout une autre fonction de test Φ' de niveau au plus α, on a

$$\mathbb{E}_\theta\left(\Phi(X_1,\cdots,X_n)\right) \geq \mathbb{E}_\theta\left(\Phi'(X_1,\cdots,X_n)\right) \quad \forall\theta\in\Theta_1.$$

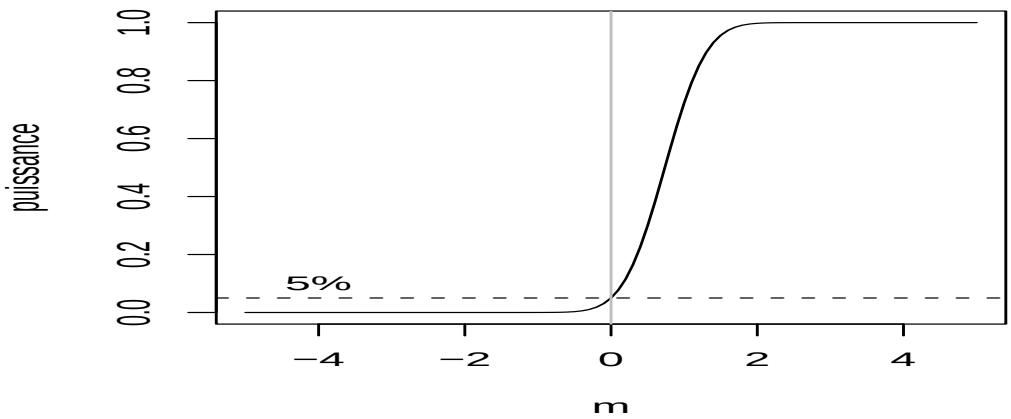


FIG. 3.1. Puissance du test $H_0 : \ m \leq m_0 \quad$ contre $\quad H_1 : \ m > m_0$,

PROPOSITION 3.2. *Le test de région critique (3.3) est UMP pour tester*

$$m = m_0 \quad \textit{contre} \quad m > m_0,$$

ou pour tester

$$m \leq m_0 \quad \textit{contre} \quad m > m_0,$$

EXERCICE 2. Donner une procédure de test pour le test suivant

$$m \geq m_0 \quad \text{contre} \quad m < m_0.$$

△△

DÉMONSTRATION. On établit ces propriétés pour le test de $m \leq m_0 \quad$ contre $\quad m > m_0$.

Tout d'abord, on a vu plus haut que $P_m\left(\overline{X}_n \geq m_0 + z_\alpha \frac{\sigma}{\sqrt{n}}\right)$ est fonction croissante de m. Il en résulte que la région critique (3.3) est sans biais et de niveau α. Ensuite, pour tout $m_1 > m_0$, elle a les propriétés du lemme de Neyman-Pearson pour tester $m = m_0$ contre $m = m_1$. Soit donc une fonction de test Φ' telle que $\mathbb{E}_m(\Phi') \leq \alpha \quad \forall m \leq m_0$. On a $\mathbb{E}_{m_0}(\Phi') \leq \alpha$, donc, d'après le lemme de Neyman-Pearson,

$$\mathbb{E}_m(\Phi') \leq P_m\left(\overline{X}_n \geq m_0 + z_\alpha \frac{\sigma}{\sqrt{n}}\right) \quad \forall m > m_0.$$

□

3.1.3. *Test bilatéral.*

DÉFINITION 3.4. *test UMPU "uniformly most powerful unbiased"*
Soit un test sans biais de niveau α pour tester Θ_0 contre Θ_1. Soit Φ sa fonction de test. On dit que ce test est uniformément le plus puissant parmi les tests sans biais

de niveau au plus α (UMPU), si pour tout une autre fonction de test Φ' sans biais de niveau au plus α, on a

$$\mathbb{E}_\theta\left(\Phi(X_1,\cdots,X_n)\right) \geq \mathbb{E}_\theta\left(\Phi'(X_1,\cdots,X_n)\right) \quad \forall\theta\in\Theta_1.$$

REMARQUE 3.2. Un test UMP est UMPU. |

Il paraît naturel, pour tester $m = m_0$ contre $m \neq m_0$, de prendre une région critique bilatérale, de la forme

$$|\overline{X}_n - m_0| \geq c.$$

On ajuste la borne c en écrivant que

$$P_{m_0}(|\overline{X}_n - m_0| \geq c) = P(|U| \geq \frac{c\sqrt{n}}{\sigma}) = \alpha,$$

où U est une variable gaussienne standard.
On en déduit que $c\sigma^{-1}\sqrt{n} = z_{\alpha/2}$. La région critique est donc

$$|\overline{X}_n - m_0| \geq z_{\alpha/2}\frac{\sigma}{\sqrt{n}}. \tag{3.4}$$

Prouvons que ce test est sans biais. Sa puissance est

$$P_m(|\overline{X}_n - m_0| \geq z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) = 1 - P(\frac{m_0-m}{\sigma}\sqrt{n} - z_{\alpha/2} \leq U \leq \frac{m_0-m}{\sigma}\sqrt{n} + z_{\alpha/2}), \tag{3.5}$$

où U est une variable gaussienne standard. Soit $h > 0$ fixé.

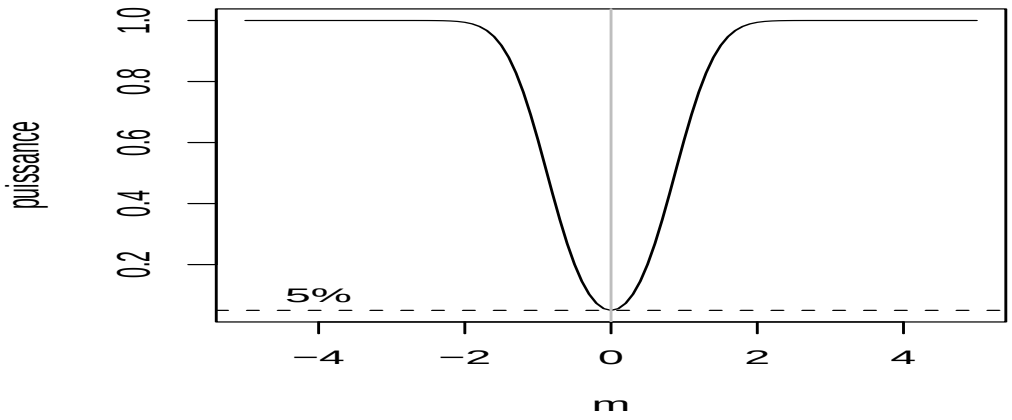


FIG. 3.2. Puissance du test $H_0 : m \leq 0$ contre $H_1 : m \neq 0$,

L'application qui à x fait correspondre

$$P(x \leq U \leq x+h) = \int_x^{x+h} \phi(t)dt$$

est dérivable et sa dérivée est

$$\phi(x+h)-\phi(x)=e^{-x^2/2}(e^{-\frac{h^2+2xh}{2}}-1).$$

on en déduit facilement que cette fonction est maximum lorsque $x=-h/2$. En posant $h=2z_{\alpha/2}$ et $x=\frac{m_0-m}{\sigma}\sqrt{n}-z_{\alpha/2}$ on conclut que la fonction de m définie en (3.5) est toujours supérieure à la valeur qu'elle prend pour $m=m_0$ qui est précisément α.
Ce test n'est pas UMP. Pour le prouver considérons la région unilatérale (3.3), qui est de niveau α. Sa puissance est $P_m(\overline{X}_n-m_0\geq z_\alpha\sigma n^{-1/2})$. D'après (3.4), la différence des puissances, exprimée en fonction de $t=\frac{m_0-m}{\sigma}\sqrt{n}$ est

$$\begin{aligned}\Delta(t) &= P_m(|\overline{X}_n-m_0|\geq z_{\alpha/2}\frac{\sigma}{\sqrt{n}})-P_m(\overline{X}_n-m_0\geq z_\alpha\frac{\sigma}{\sqrt{n}})\\ &= P(U\geq z_{\alpha/2}+t)+P(U\leq -z_{\alpha/2}+t)-P(U\geq z_\alpha+t).\end{aligned}$$

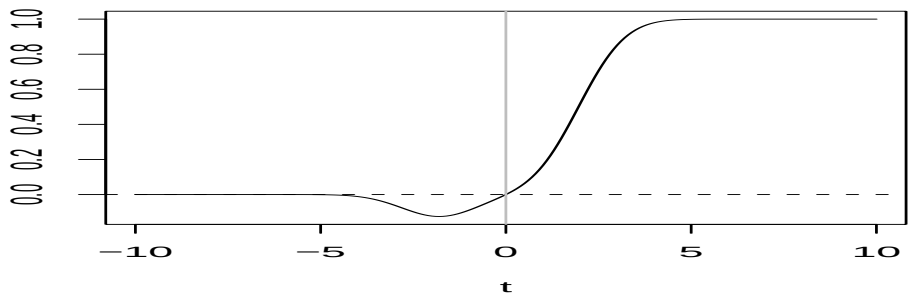


FIG. 3.3. Fonction $t\to\Delta(t)$

Cette fonction est nulle en $t=0$ et tend vers zéro quand $t\to-\infty$. Sa dérivée est

$$\begin{aligned}&\Delta'(t)=-\phi(z_{\alpha/2}+t)+\phi(-z_{\alpha/2}+t)+\phi(z_\alpha+t)\\ =\ &\frac{\exp\left(-\frac{(z_{\alpha/2}+t)^2}{2}\right)}{\sqrt{2}\pi}\left(\exp(2tz_{\alpha/2})-1+\exp\frac{-2t(z_\alpha-z_{\alpha/2})-z_\alpha^2+z_{\alpha/2}^2}{2}\right).\end{aligned}$$

Il est facile de voir que la fonction

$$t\mapsto\exp(2tz_{\alpha/2})-1+\exp\frac{-2t(z_\alpha-z_{\alpha/2})-z_\alpha^2+z_{\alpha/2}^2}{2}$$

est strictement croissante. De plus, elle varie de -1 à $\exp\frac{-z_\alpha^2+z_{\alpha/2}^2}{2}>0$ lorsque t croit de $-\infty$ à 0. On en déduit que la fonction $\Delta(t)$ est strictement négative pour $t<0$. La puissance du test unilatéral est donc strictement supérieure celle du test bilatéral pour $m>m_0$, et le test bilatéral n'est pas UMP.

Cependant, on remarque que la région critique unilatérale n'est pas sans biais. En effet la puissance

$$P_m\left(\overline{X}_n \geq m_0 + z_\alpha \frac{\sigma}{\sqrt{n}}\right) \quad m \neq m_0$$

étant fonction croissante de m, elle est inférieure à α pour $m < m_0$. On démontre (on ne le fera pas ici) que le test bilatéral (3.4) est UMPU.

3.2. Tests de comparaison à une valeur donnée de l'espérance d'un échantillon gaussien de variance inconnue. Bien qu'on ne connaisse pas la variance, on ne fait pas ici porter le test sur ce paramètre. On dit parfois qu'on le traite comme un paramètre parasite.
On désire tester, comme précédemment, $m \leq m_0$ contre $m > m_0$ ou bien $m = m_0$ contre $m \neq m_0$.
La première remarque à faire est que les tests précédents ne sont plus utilisables puisque leur région critique utilise le paramètre inconnu σ^2. Une idée s'impose : remplacer σ^2 par un estimateur.
Commençons par rappeler que puisque les X_j sont indépendantes et toutes de loi gaussienne $\mathcal{N}(m, \sigma^2)$, les deux statistiques

$$\overline{X}_n \quad \text{et} \quad \hat{s}_n^2 = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \overline{X}_n)^2$$

sont indépendantes. D'autre part, $\frac{n-1}{\sigma^2}\hat{s}_n^2$ est distribué selon une loi du χ^2 à $n-1$ degrés de libertés. Clairement, la statistique qui va nous intéresser est le rapport

$$T_{n-1} = \frac{\sqrt{n}(\overline{X}_n - m)}{\hat{s}_n}$$

qui est obtenu à partir de

$$\frac{\sqrt{n}(\overline{X}_n - m)}{\sigma}$$

en remplaçant σ par son estimateur. Cette statistique T_{n-1} est distribué selon une loi de Student à $n-1$ degrés de liberté. Cette loi a une densité paire. Elle est tabulée. Pour tout $\alpha \in]0,1[$ on définira, comme pour la loi gaussienne, $t_{n-1,\alpha}$ par

$$P(T_{n-1} \geq t_{n-1,\alpha}) = \alpha.$$

Tenant compte de ce qui précède, on prendra comme région critique pour tester $m \leq m_0$ contre $m > m_0$

$$\overline{X}_n \geq m_0 + t_{n-1,\alpha}\frac{\hat{s}_n}{\sqrt{n}}$$

et, pour tester $m = m_0$ contre $m \neq m_0$

$$|\overline{X}_n - m_0| \geq t_{n-1,\alpha/2}\frac{\hat{s}_n}{\sqrt{n}}.$$

Pour terminer, on remarque que

$$T_{n-1} = \left(\frac{\sqrt{n}(\overline{X}_n - m)}{\sigma}\right)\left(\frac{\sigma}{\hat{s}_n}\right).$$

Comme, lorsque $n \to \infty$, le deuxième facteur tend presque sûrement vers 1 tandis que le premier converge en loi vers une gaussienne standard, on conclut que T_{n-1} converge en loi vers une gaussienne standard. C'est d'ailleurs pour cette raison que la loi de Student n'est pas tabulée pour les grandes valeurs du degré de liberté.

Dans ce cas on utilise directement la loi gaussienne standard.

3.3. Tests de comparaison des espérances de deux échantillons gaussiens. On considère deux échantillons gaussiens $(X_1, \cdots, X_{n_1})$ et $(Y_1, \cdots, Y_{n_2})$. On a coutume de distinguer deux cas. Ou les échantillons sont indépendants (il s'agira par exemple de comparer les tailles moyennes des individus de deux populations), ou ils sont appariés (c'est le cas dans l'exemple 1.2 du début de ce cours). Dans le second cas, évidemment $n_1 = n_2$.
Le cadre est, dans les deux cas, le suivant : les X_j sont indépendantes, toutes de loi $\mathcal{N}(m_1, \sigma_1^2)$ et les Y_j sont indépendantes, toutes de loi $\mathcal{N}(m_2, \sigma_2^2)$. On désire tester

$$m_1 = m_2 \quad \text{contre} \quad m_1 \neq m_2.$$

(Le lecteur adaptera lui-même les résultats aux tests unilatéraux de $m_1 \leq m_2$ contre $m_1 > m_2$ et de $m_1 \geq m_2$ contre $m_1 < m_2$).

3.3.1. *Échantillons indépendants, variances connues.* Il est naturel de baser le test sur la différence des moyennes arithmétiques, donc de prendre une région critique de la forme

$$|\overline{X}_{n_1} - \overline{Y}_{n_2}| \geq c.$$

On ajuste c en écrivant que

$$P_{m_1=m_2}(|\overline{X}_{n_1} - \overline{Y}_{n_2}| \geq c) = \alpha. \tag{3.6}$$

Pour calculer c on remarque que

$$U := \frac{\overline{X}_{n_1} - \overline{Y}_{n_2} - (m_1 - m_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

a une loi gaussienne standard. En conséquence,

$$c = z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}},$$

et la région critique est

$$|\overline{X}_{n_1} - \overline{Y}_{n_2}| \geq z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

Montrons que ce test est sans biais. Sa puissance est égale à

$$\begin{aligned}
& P_{m_1,m_2}\left(\overline{X}_{n_1}-\overline{Y}_{n_2}\geq z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\right)+P_{m_1,m_2}\left(\overline{X}_{n_1}-\overline{Y}_{n_2}\leq -z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\right)\\
= \quad & P_{m_1,m_2}\left(\frac{\overline{X}_{n_1}-\overline{Y}_{n_2}-(m_1-m_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}\geq z_{\alpha/2}-\frac{(m_1-m_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}\right)\\
& \qquad +P_{m_1,m_2}\left(\frac{\overline{X}_{n_1}-\overline{Y}_{n_2}-(m_1-m_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}\leq -z_{\alpha/2}-\frac{(m_1-m_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}\right)\\
= \quad & P\left(U\geq z_{\alpha/2}-\frac{(m_1-m_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}\right)+P\left(U\leq -z_{\alpha/2}-\frac{(m_1-m_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}\right).
\end{aligned}$$

On a vu plus haut que $P\left(U\geq z_{\alpha/2}-t\right)+P\left(U\leq -z_{\alpha/2}-t\right)$ est toujours supérieur à α, valeur atteinte pour $t=0$. Le test est donc sans biais.

3.3.2. *Échantillons indépendants, variances inconnues mais égales.* Il se trouve que, si les variances σ_1^2 et σ_2^2 sont inconnues, on ne sait construire une procédure de niveau α exacte que si $\sigma_1^2=\sigma_2^2=:\sigma^2$, ce que l'on supposera ici.
On utilise, pour construire le test, le fait que la somme de deux variables indépendantes distribuées selon des lois du χ^2 est encore distribuée selon une loi du χ^2 et les degrés de libertés s'ajoutent. En conséquence

$$\frac{\sum_{j=1}^n(X_j-\overline{X}_{n_1})^2+\sum_{j=1}^n(Y_j-\overline{Y}_{n_2})^2}{\sigma^2}$$

a une loi du χ^2 à n_1+n_2-2 degrés de liberté. De plus, cette statistique est indépendante de $\overline{X}_{n_1}-\overline{Y}_{n_2}$, si bien que, sous l'hypothèse nulle $m_1=m_2$, la statistique

$$\begin{aligned}
& \frac{\overline{X}_{n_1}-\overline{Y}_{n_2}}{\sqrt{\sigma^2(1/n_1+1/n_2)}}\left(\frac{\sum_{j=1}^n(X_j-\overline{X}_{n_1})^2+\sum_{j=1}^n(Y_j-\overline{Y}_{n_2})^2}{\sigma^2(n_1+n_2-2)}\right)^{-1/2}\\
= \quad & \frac{\overline{X}_{n_1}-\overline{Y}_{n_2}}{\sqrt{\sum_{j=1}^n(X_j-\overline{X}_{n_1})^2+\sum_{j=1}^n(Y_j-\overline{Y}_{n_2})^2}}\left(\frac{n_1+n_2-2}{1/n_1+1/n_2}\right)^{1/2}
\end{aligned}$$

a une loi de Student à n_1+n_2-2 degrés de liberté. On prendra donc une région critique de la forme

$$\frac{|\overline{X}_{n_1}-\overline{Y}_{n_2}|}{\sqrt{\sum_{j=1}^n(X_j-\overline{X}_{n_1})^2+\sum_{j=1}^n(Y_j-\overline{Y}_{n_2})^2}}\geq c,$$

et on ajuste c en s'aidant de la table de la loi de Student à n_1+n_2-2 degrés de liberté ce qui donne $c=t_{n_1+n_2-2,\alpha/2}\left(\frac{n_1+n_2-2}{1/n_1+1/n_2}\right)^{-1/2}$, d'où la région critique

$$\frac{|\overline{X}_{n_1}-\overline{Y}_{n_2}|}{\sqrt{\sum_{j=1}^n(X_j-\overline{X}_{n_1})^2+\sum_{j=1}^n(Y_j-\overline{Y}_{n_2})^2}}\geq t_{n_1+n_2-2,\alpha/2}\left(\frac{1/n_1+1/n_2}{n_1+n_2-2}\right)^{1/2}.$$

3.3.3. *Échantillons appariés.* On suppose que les vecteurs aléatoires $(X_1, Y_1), \cdots, (X_n, Y_n)$ sont mutuellement indépendants et tous de même loi gaussienne d'espérance ${}^t(m_1, m_2)$. On ne connaît pas la matrice de covariance et on désire tester l'hypothèse nulle $m_1 = m_2$ contre l'alternative $m_1 \neq m_2$. (Au lecteur d'adapter les résultats aux tests unilatéraux).
Le test sera basé sur les différences $\Delta_j = X_j - Y_j$. Les Δ_j sont des variables gaussiennes indépendantes et toutes de même loi. Sous l'hypothèse nulle elles sont de plus centrées. On est donc ramenés au test de comparaison à zéro de l'espérance d'un échantillon gaussien dont la variance n'est pas connue. Posons $\overline{\Delta}_n = \overline{X}_n - \overline{Y}_n$ et $\hat{s}_n^2 = \frac{1}{n-1}\sum_{j=1}^n (\Delta_j - \overline{\Delta}_n)^2$. La région critique de niveau α pour le test bilatéral s'écrit alors

$$|\overline{\Delta}_n| \geq t_{n-1,\alpha/2}\frac{\hat{s}_n}{\sqrt{n}}.$$

4. Tests sur les espérances d'échantillons non gaussiens

Traitons par exemple le test de $m = m_0$ contre $m > m_0$. Ce test (comme la plupart de ceux qui viennent d'être abordés) est basé sur la moyenne arithmétique. Si les variables X_j sont indépendantes, toutes de même loi et si $\mathbb{E}(X_j^2) < \infty$, le théorème central limite permet d'affirmer que, si l'espérance est m, $(\overline{X}_n - m)\frac{\sqrt{n}}{\sigma}$ converge en loi vers une variable gaussienne standard U. Plus exactement

$$P_{m_0}(\overline{X}_n \geq m_0 + z_\alpha \frac{\sigma}{\sqrt{n}}) - P\left(U \geq z_\alpha\right) \to 0, \quad \text{quand} \quad n \to \infty,$$

ce qui signifie que le niveau du test tend vers α quand $n \to \infty$. Quant à la puissance du test on a pour $m > m_0$

$$P_m\left(\overline{X}_n \geq m_0 + z_\alpha \frac{\sigma}{\sqrt{n}}\right) - P\left(U \geq \sqrt{n}\frac{m_0 - m}{\sigma} + z_\alpha\right) \to 0, \quad \text{quand} \quad n \to \infty.$$

La puissance est donc équivalente à ce qu'elle est sous l'hypothèse gaussienne.
On laisse au lecteur le soin d'inspecter les autres tests des sections précédentes.

CHAPITRE 4

Tables statistiques

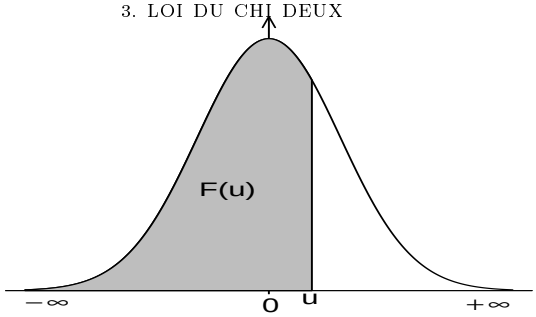
Ces tables ont été construites à l'aide des fonctions quantiles et des fonctions de répartitions [1] disponibles dans le logiciel R version 1.6.1 décembre 2002.

1. Loi Gaussienne

2. Loi de Student

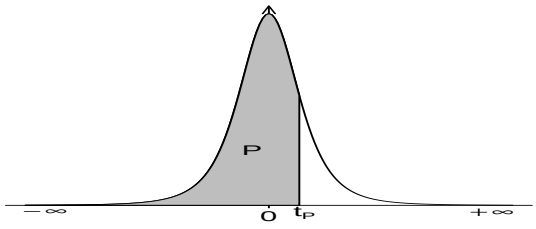
3. Loi du Chi deux

[1] pnorm,qchisq,qt



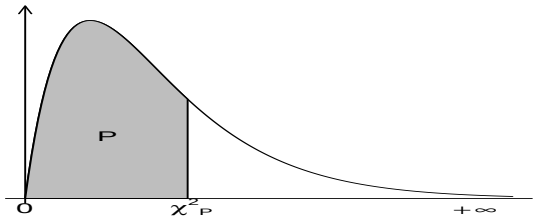
TAB. 1. Fonction de répartion F de la loi normale standard $X \sim \mathcal{N}(0,1)$. La table ci-dessous donne la valeur $F(u) = P(X \leq u)$ en fonction de u. Par exemple si $u = 1.96 = 1.9 + 0.06$ alors $F(u) = 0.975$

$u = u_1 + u_2$ (u_1 \ u_2)	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0	0.5	0.5039	0.5079	0.5119	0.5159	0.5199	0.5239	0.5279	0.5318	0.5358
0.1	0.5398	0.5437	0.5477	0.5517	0.5556	0.5596	0.5635	0.5674	0.5714	0.5753
0.2	0.5792	0.5831	0.587	0.5909	0.5948	0.5987	0.6025	0.6064	0.6102	0.614
0.3	0.6179	0.6217	0.6255	0.6293	0.633	0.6368	0.6405	0.6443	0.648	0.6517
0.4	0.6554	0.659	0.6627	0.6664	0.67	0.6736	0.6772	0.6808	0.6843	0.6879
0.5	0.6914	0.6949	0.6984	0.7019	0.7054	0.7088	0.7122	0.7156	0.719	0.7224
0.6	0.7257	0.729	0.7323	0.7356	0.7389	0.7421	0.7453	0.7485	0.7517	0.7549
0.7	0.758	0.7611	0.7642	0.7673	0.7703	0.7733	0.7763	0.7793	0.7823	0.7852
0.8	0.7881	0.791	0.7938	0.7967	0.7995	0.8023	0.8051	0.8078	0.8105	0.8132
0.9	0.8159	0.8185	0.8212	0.8238	0.8263	0.8289	0.8314	0.8339	0.8364	0.8389
1	0.8413	0.8437	0.8461	0.8484	0.8508	0.8531	0.8554	0.8576	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8707	0.8728	0.8749	0.8769	0.8789	0.8809	0.8829
1.2	0.8849	0.8868	0.8887	0.8906	0.8925	0.8943	0.8961	0.8979	0.8997	0.9014
1.3	0.9031	0.9049	0.9065	0.9082	0.9098	0.9114	0.913	0.9146	0.9162	0.9177
1.4	0.9192	0.9207	0.9221	0.9236	0.925	0.9264	0.9278	0.9292	0.9305	0.9318
1.5	0.9331	0.9344	0.9357	0.9369	0.9382	0.9394	0.9406	0.9417	0.9429	0.944
1.6	0.9452	0.9463	0.9473	0.9484	0.9494	0.9505	0.9515	0.9525	0.9535	0.9544
1.7	0.9554	0.9563	0.9572	0.9581	0.959	0.9599	0.9607	0.9616	0.9624	0.9632
1.8	0.964	0.9648	0.9656	0.9663	0.9671	0.9678	0.9685	0.9692	0.9699	0.9706
1.9	0.9712	0.9719	0.9725	0.9731	0.9738	0.9744	0.975	0.9755	0.9761	0.9767
2	0.9772	0.9777	0.9783	0.9788	0.9793	0.9798	0.9803	0.9807	0.9812	0.9816
2.1	0.9821	0.9825	0.9829	0.9834	0.9838	0.9842	0.9846	0.9849	0.9853	0.9857
2.2	0.986	0.9864	0.9867	0.9871	0.9874	0.9877	0.988	0.9883	0.9886	0.9889
2.3	0.9892	0.9895	0.9898	0.99	0.9903	0.9906	0.9908	0.9911	0.9913	0.9915
2.4	0.9918	0.992	0.9922	0.9924	0.9926	0.9928	0.993	0.9932	0.9934	0.9936
2.5	0.9937	0.9939	0.9941	0.9942	0.9944	0.9946	0.9947	0.9949	0.995	0.9952
2.6	0.9953	0.9954	0.9956	0.9957	0.9958	0.9959	0.996	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.997	0.9971	0.9971	0.9972	0.9973
2.8	0.9974	0.9975	0.9975	0.9976	0.9977	0.9978	0.9978	0.9979	0.998	0.998
2.9	0.9981	0.9981	0.9982	0.9983	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986



TAB. 2. La table ci-dessous donne les quantiles t_P de la loi de Student en fonction de P et ν le nombre de degrés de liberté. Si $X \sim \mathcal{T}(\nu)$ alors $P = P(X \leq t_P)$. Par exemple si X suit une loi de Student à $\nu = 8$ degrés de liberté alors pour $P = .95$ on obtient $t_P = 1.859$

ν \ P	0.55	0.6	0.65	0.7	0.75	0.8	0.85	0.9	0.95	0.975	0.99
1	0.158	0.324	0.509	0.726	1	1.376	1.962	3.077	6.313	12.706	31.82
2	0.142	0.288	0.444	0.617	0.816	1.06	1.386	1.885	2.919	4.302	6.964
3	0.136	0.276	0.424	0.584	0.764	0.978	1.249	1.637	2.353	3.182	4.54
4	0.133	0.27	0.414	0.568	0.74	0.94	1.189	1.533	2.131	2.776	3.746
5	0.132	0.267	0.408	0.559	0.726	0.919	1.155	1.475	2.015	2.57	3.364
6	0.131	0.264	0.404	0.553	0.717	0.905	1.134	1.439	1.943	2.446	3.142
7	0.13	0.263	0.401	0.549	0.711	0.896	1.119	1.414	1.894	2.364	2.997
8	0.129	0.261	0.399	0.545	0.706	0.888	1.108	1.396	1.859	2.306	2.896
9	0.129	0.26	0.397	0.543	0.702	0.883	1.099	1.383	1.833	2.262	2.821
10	0.128	0.26	0.396	0.541	0.699	0.879	1.093	1.372	1.812	2.228	2.763
11	0.128	0.259	0.395	0.539	0.697	0.875	1.087	1.363	1.795	2.2	2.718
12	0.128	0.259	0.394	0.538	0.695	0.872	1.083	1.356	1.782	2.178	2.68
13	0.128	0.258	0.393	0.537	0.693	0.87	1.079	1.35	1.77	2.16	2.65
14	0.127	0.258	0.393	0.536	0.692	0.868	1.076	1.345	1.761	2.144	2.624
15	0.127	0.257	0.392	0.535	0.691	0.866	1.073	1.34	1.753	2.131	2.602
16	0.127	0.257	0.392	0.535	0.69	0.864	1.071	1.336	1.745	2.119	2.583
17	0.127	0.257	0.391	0.534	0.689	0.863	1.069	1.333	1.739	2.109	2.566
18	0.127	0.257	0.391	0.533	0.688	0.862	1.067	1.33	1.734	2.1	2.552
19	0.127	0.256	0.391	0.533	0.687	0.86	1.065	1.327	1.729	2.093	2.539
20	0.127	0.256	0.39	0.532	0.686	0.859	1.064	1.325	1.724	2.085	2.527
21	0.127	0.256	0.39	0.532	0.686	0.859	1.062	1.323	1.72	2.079	2.517
22	0.127	0.256	0.39	0.532	0.685	0.858	1.061	1.321	1.717	2.073	2.508
23	0.127	0.256	0.39	0.531	0.685	0.857	1.06	1.319	1.713	2.068	2.499
24	0.126	0.256	0.389	0.531	0.684	0.856	1.059	1.317	1.71	2.063	2.492
25	0.126	0.256	0.389	0.531	0.684	0.856	1.058	1.316	1.708	2.059	2.485
26	0.126	0.255	0.389	0.53	0.684	0.855	1.057	1.314	1.705	2.055	2.478
27	0.126	0.255	0.389	0.53	0.683	0.855	1.056	1.313	1.703	2.051	2.472
28	0.126	0.255	0.389	0.53	0.683	0.854	1.055	1.312	1.701	2.048	2.467
29	0.126	0.255	0.389	0.53	0.683	0.854	1.055	1.311	1.699	2.045	2.462
30	0.126	0.255	0.389	0.53	0.682	0.853	1.054	1.31	1.697	2.042	2.457



TAB. 3. La table ci-dessous donne les quantiles χ^2_P de la loi du χ^2 en fonction de P et ν le nombre de degrés de liberté. Si $X \sim \chi^2(\nu)$ alors $P = P(X \leq \chi^2_P)$. Par exemple si X suit une loi du χ^2 à $\nu = 5$ degrés de liberté alors pour $P = .95$ on obtient $\chi^2_P = 11.07$

ν \ P	0.01	0.025	0.05	0.1	0.25	0.5	0.75	0.9	0.95	0.975	0.99	0.995
1	0.0002	0.001	0.0039	0.01	0.1	0.45	1.32	2.7	3.84	5.02	6.63	7.87
2	0.02	0.05	0.1	0.21	0.57	1.38	2.77	4.6	5.99	7.37	9.21	10.59
3	0.11	0.21	0.35	0.58	1.21	2.36	4.1	6.25	7.81	9.34	11.34	12.83
4	0.29	0.48	0.71	1.06	1.92	3.35	5.38	7.77	9.48	11.14	13.27	14.86
5	0.55	0.83	1.14	1.61	2.67	4.35	6.62	9.23	11.07	12.83	15.08	16.74
6	0.87	1.23	1.63	2.2	3.45	5.34	7.84	10.64	12.59	14.44	16.81	18.54
7	1.23	1.68	2.16	2.83	4.25	6.34	9.03	12.01	14.06	16.01	18.47	20.27
8	1.64	2.17	2.73	3.48	5.07	7.34	10.21	13.36	15.5	17.53	20.09	21.95
9	2.08	2.7	3.32	4.16	5.89	8.34	11.38	14.68	16.91	19.02	21.66	23.58
10	2.55	3.24	3.94	4.86	6.73	9.34	12.54	15.98	18.3	20.48	23.2	25.18
11	3.05	3.81	4.57	5.57	7.58	10.34	13.7	17.27	19.67	21.92	24.72	26.75
12	3.57	4.4	5.22	6.3	8.43	11.34	14.84	18.54	21.02	23.33	26.21	28.29
13	4.1	5	5.89	7.04	9.29	12.33	15.98	19.81	22.36	24.73	27.68	29.81
14	4.66	5.62	6.57	7.78	10.16	13.33	17.11	21.06	23.68	26.11	29.14	31.31
15	5.22	6.26	7.26	8.54	11.03	14.33	18.24	22.3	24.99	27.48	30.57	32.8
16	5.81	6.9	7.96	9.31	11.91	15.33	19.36	23.54	26.29	28.84	31.99	34.26
17	6.4	7.56	8.67	10.08	12.79	16.33	20.48	24.76	27.58	30.19	33.4	35.71
18	7.01	8.23	9.39	10.86	13.67	17.33	21.6	25.98	28.86	31.52	34.8	37.15
19	7.63	8.9	10.11	11.65	14.56	18.33	22.71	27.2	30.14	32.85	36.19	38.58
20	8.26	9.59	10.85	12.44	15.45	19.33	23.82	28.41	31.41	34.16	37.56	39.99
21	8.89	10.28	11.59	13.23	16.34	20.33	24.93	29.61	32.67	35.47	38.93	41.4
22	9.54	10.98	12.33	14.04	17.23	21.33	26.03	30.81	33.92	36.78	40.28	42.79
23	10.19	11.68	13.09	14.84	18.13	22.33	27.14	32	35.17	38.07	41.63	44.18
24	10.85	12.4	13.84	15.65	19.03	23.33	28.24	33.19	36.41	39.36	42.97	45.55
25	11.52	13.11	14.61	16.47	19.93	24.33	29.33	34.38	37.65	40.64	44.31	46.92
26	12.19	13.84	15.37	17.29	20.84	25.33	30.43	35.56	38.88	41.92	45.64	48.28
27	12.87	14.57	16.15	18.11	21.74	26.33	31.52	36.74	40.11	43.19	46.96	49.64
28	13.56	15.3	16.92	18.93	22.65	27.33	32.62	37.91	41.33	44.46	48.27	50.99
29	14.25	16.04	17.7	19.76	23.56	28.33	33.71	39.08	42.55	45.72	49.58	52.33
30	14.95	16.79	18.49	20.59	24.47	29.33	34.79	40.25	43.77	46.97	50.89	53.67

CHAPITRE 5

Tests du χ^2

On aborde ici quelques exemples de tests d'ajustement à une loi donnée ou à une famille donnée de lois. On observe $X_1, \cdots, X_n$ un échantillon d'une loi P, et on veut tester soit que $P = P_0$ soit que P fait partie d'une famille paramétrée de lois $(P_\theta)_{\theta \in \Theta}$.

1. Tests d'ajustement à une loi discrète donnée

La loi des X_j est concentrée sur l'ensemble fini $\{1, \cdots, m\}$. Autrement dit on a, avec $p_j := P(X_1 = j) \quad j = 1, \cdots, m$

$$p_j > 0 \quad \forall j \qquad \text{et} \qquad \sum_{j=1}^{m} p_j = 1.$$

Notons

$$N_h^{(n)} = \sum_{j=1}^{n} \mathbb{I}_{\{h\}}(X_j) \quad \forall h \in \{1, \cdots, m\} \quad \text{et} \quad N_n = {}^t(N_1^{(n)}, \cdots, N_m^{(n)}).$$

On a

$$\mathbb{E}_\Pi N_n = {}^t(p_1, \cdots, p_m).$$

On désire tester l'hypothèse $\Pi = \Pi_0 = {}^t(p_1^0, \cdots, p_m^0)$. On base le test sur la statistique

$$Q_n = \sum_{j=1}^{m} \frac{(N_j^{(n)} - np_j^0)^2}{np_j^0},$$

qui mesure l'écart entre les effectifs observés $N_j^{(n)}$ et leurs espérances sous l'hypothèse nulle. On décide de rejeter l'hypothèse $\Pi = \Pi_0$ si $Q_n \geq c$ et de l'accepter dans le cas contraire. Pour ajuster c on devrait connaître la loi de Q_n sous H_0. Malheureusement il n'en est rien, on ne connaît que sa loi limite qui ne permettra que la construction d'une région critique de niveau asymptotique α. L'étude de la loi limite de Q_n est basée sur la proposition suivante :

PROPOSITION 5.1. *Lorsque $n \to \infty$, la suite de vecteurs aléatoires $n^{-1/2}(N_n - n\Pi)$ converge en loi vers un vecteur V gaussien centré de matrice de covariance*

$$\begin{pmatrix} p_1(1-p_1) & -p_1p_2 & \cdots & -p_1p_m \\ -p_1p_2 & p_2(1-p_2) & \cdots & -p_2p_m \\ \vdots & \vdots & \vdots & \vdots \\ -p_mp_1 & -p_mp_2 & \cdots & p_m(1-p_m) \end{pmatrix}. \tag{5.1}$$

DÉMONSTRATION. Ce résultat n'est autre que le théorème central limite dans $\mathbb{R}^m$ appliqué au vecteur multinômial N_n En effet, le vecteur N_n s'écrit comme la somme de vecteurs aléatoires de $\mathcal{L}^2$, indépendants et de même loi. On a

$$N_n = \sum_{k=1}^{n} {}^t(\mathbb{I}_{\{1\}}(X_k), \dots, \mathbb{I}_{\{m\}}(X_k))$$

avec

$$\begin{aligned}
&\mathbb{E}^t(\mathbb{I}_{\{1\}}(X_1), \dots, \mathbb{I}_{\{m\}}(X_1)) = {}^t(p_1, \dots, p_m)\\
&\mathrm{Var}(\mathbb{I}_{\{j\}}(X_1)) = p_j(1-p_j) \qquad j \in \{1, \dots, m\}\\
&\mathbb{E}\,\mathbb{I}_{\{j\}}(X_1)\mathbb{I}_{\{k\}}(X_1) = 0 \qquad j \neq k,\ j,k \in \{1, \dots, m\}
\end{aligned}$$

(5.1) est donc la matrice de covariance des vecteurs aléatoires ${}^t(\mathbb{I}_{\{1\}}(X_k), \dots, \mathbb{I}_{\{m\}}(X_k))$. Le résultat découle ensuite du Théorème central limite.

□

Cette normalité asymptotique implique à que la loi limite de Q_n ets une loi du χ^2.

THÉORÈME 5.2. *Lorsque $n \to \infty$, Q_n converge en loi vers une variable Q dont la loi est la loi du χ^2 à $m-1$ degrés de libertés.*

DÉMONSTRATION. Soit

$$f(v) = \sum_{j=1}^{m} \frac{v_j^2}{p_j^0} \qquad \forall v = {}^t(v_1, \cdots, v_m) \in \mathbb{R}^m.$$

On a évidemment $Q_n = f\left(n^{-1/2}(N_n - n\Pi_0)\right)$. La fonction f est continue, si bien que, d'après la proposition 5.1, Q_n converge en loi vers $f(V)$. Notons

$$W = {}^t\left(\frac{V_1}{\sqrt{p_1^0}}, \cdots, \frac{V_m}{\sqrt{p_m^0}}\right).$$

Le vecteur W est gaussien, centré. D'après la proposition 5.1 sa matrice de covariances est

$$\Sigma_W = I - \Pi_0^{1/2}\,{}^t\Pi_0^{1/2} \quad \text{avec} \quad {}^t\Pi_0^{1/2} = (\sqrt{p_1^0}, \cdots, \sqrt{p_m^0}),$$

et on a $f(V) = \|W\|_2{}^2$.
Considérons alors une matrice orthonormée A telle que $A\Pi_0^{1/2} = {}^t(0, \cdots, 0, 1)$.
On a à la fois

$$\Sigma_{AW} = A\Sigma_W{}^tA = I - A\Pi_0^{1/2}\,{}^t\Pi_0^{1/2}\,{}^tA = \begin{pmatrix} 1 & 0 & \dots & 0\\ 0 & 1 & \dots & 0\\ \vdots & \vdots & \ddots & 0\\ 0 & \dots & 0 & 1\\ 0 & 0 & 0 & 0 \end{pmatrix},$$

et $\|AW\|_2{}^2 = \|W\|_2{}^2$, d'où le résultat annoncé puisque $\|AW\|_2{}^2$ a une loi du χ^2 à $m-1$ degrés de liberté. □

La construction du test s'ensuit évidemment. Soit Q une variable aléatoire distribuée selon une loi du χ^2 à k degrés de libertés et notons $\chi_{k,\alpha}$ la valeur de x telle que $P(Q > \chi_{k,\alpha}) = \alpha$. D'après ce qui précède la région critique

$$Q_n > \chi_{m-1,\alpha}$$

a un niveau qui tend vers α lorsque $n \to \infty$.
Voici un exemple d'application. On désire tester l'hypothèse H_0 : la loi P_0 est une loi gaussienne d'espérance m et de variance σ^2, paramètres supposés connus. On divise la droite en m intervalles

$$I_1 =]-\infty, x_1], I_2 = [x_1, x_2], \cdots, I_m =]x_{m-1}, +\infty[\tag{5.2}$$

et on calcule les probabilités $p_j = P_0(X_1 \in I_j)$. D'après la proposition 5.2, la statistique

$$Q_n = \sum_{j=1}^{m} \frac{(N_j^{(n)} - np_j)^2}{np_j}$$

converge en loi vers une variable distribuée selon une loi du χ^2 à $m-1$ degrés de liberté.

Remarque importante : le regroupement en intervalles (ou classes) n'est pas indispensable. On peut avoir recours au test de *Kolmogorov-Smirnov*. Ce test utilise la loi limite de

$$D_n := \sqrt{n} \sup_x |\hat{F}_n(x) - F(x)| \tag{5.3}$$

où

$$\hat{F}_n(x) = \frac{1}{n} \sum_{j=1}^{n} \mathbb{I}_{]-\infty,x]}(X_j)$$

est la fonction de répartition empirique et où F est la fonction de répartition de la loi testée (ici, la loi $\mathcal{N}(m, \sigma^2)$). La loi limite de D_n est tabulée et en outre D_n est facile à calculer car le sup dans (5.3) est atteint en un des points $x = X_j$ (pour s'en persuader, il suffit de tracer les graphes des deux fonctions F et $\hat{F}_n$).

2. Tests d'ajustement à une famille paramétrée de lois

Les lois sont toutes concentrées sur le même ensemble discret $\{1, \cdots, m\}$. On désire tester que la loi commune des X_j fait partie de la famille P_θ ($\theta \in \Theta$, ouvert de $\mathbb{R}^p$), définie par

$$p_j(\theta) = P_\theta(X_1 = j),\ \forall j \in \{1, \cdots, m\}, \quad \text{où} \quad \theta \in \Theta, \quad \text{ouvert de } \mathbb{R}^p.$$

La première idée qui vient à l'esprit est d'utiliser la statistique Q_n du paragraphe précédent après avoir remplacé θ par un estimateur $\hat{\theta}_n$ dans l'expression des $p_j(\theta)$. On admettra sans démonstration la proposition suivante qui prouve que la statistique ainsi modifiée a encore pour loi limite une loi du χ^2, mais avec un degré de liberté plus petit.

PROPOSITION 5.3. *Supposons $p < m$. Soit $\hat{\theta}_n$ une suite d'estimateurs du maximum de vraisemblance de θ vérifiant les conditions de convergence et de normalité asymptotique des théorèmes 2.3 et 2.4. La suite*

$$\hat{Q}_n = \sum_{j=1}^{m} \frac{(N_j^{(n)} - np_j(\hat{\theta}_n))^2}{np_j(\hat{\theta}_n)}$$

converge en loi vers Q, variable distribuée selon une loi du χ^2 à $m-p-1$ degrés de liberté.

On remarque que c'est le nombre de paramètres estimés qui compte dans la diminution du degré de libertés de la loi du χ^2 limite.
Voici un exemple d'application. On désire tester l'hypothèse H_0 : la loi P est une loi gaussienne de variance 1. Autrement dit la famille P_θ est la famille des lois $\mathcal{N}(\theta,1)$. On estime θ par la moyenne arithmétique $\overline{X}_n$. On reprend le découpage de la droite (5.2) et on estime les probabilités $p_j = P_\theta(X_j \in I_j)$ par $\hat{p}_j = P_{\overline{X}_n}(X_j \in I_j)$. D'après la proposition 5.3, la statistique

$$\hat{Q}_n = \sum_{j=1}^{m} \frac{(N_j^{(n)} - n\hat{p}_j)^2}{n\hat{p}_j}$$

converge en loi vers une variable distribuée selon une loi du χ^2 à $m-2$ degrés de liberté.

3. Test d'indépendance

Considérons n vecteurs aléatoires i.i.d. $(X_1,Y_1),\cdots,(X_n,Y_n)$ à valeurs dans l'ensemble fini $E = \{1,\cdots,s\}\times\{1,\cdots,t\}$. On désire tester l'indépendance des deux coordonnées X_1 et Y_1. Notant

$$P(X_1=i,Y_1=j) = p_{i,j} \quad \text{et} \quad p_{i,\cdot} = \sum_{j=1}^{t} p_{i,j}, \quad p_{\cdot,j} = \sum_{i=1}^{s} p_{i,j},$$

on veut tester que la loi sur l'ensemble E caractérisée par les $p_{i,j}$ est en fait la loi produit

$$p_{i,j} = p_{i,\cdot}p_{\cdot,j} \quad \forall (i,j)\in E. \tag{5.4}$$

Pour $(i,j)\in E$, notons $N_{i,j}$ le nombre de couples (X_k,Y_k) tels que $X_k=i$ et $Y_k=j$ et notons

$$N_{i,\cdot} = \sum_{j=1}^{t} N_{i,j}, \quad N_{\cdot,j} = \sum_{i=1}^{s} N_{i,j}.$$

Proposition 5.4. *Sous l'hypothèse d'indépendance, la statistique*

$$\hat{Q}_n = \sum_{(i,j)\in E} \frac{\left(N_{i,j} - \frac{N_{i,\cdot}N_{\cdot,j}}{n}\right)^2}{\frac{N_{i,\cdot}N_{\cdot,j}}{n}},$$

converge en loi, vers une variable ayant une loi du χ^2 à $(s-1)(t-1)$ degrés de liberté.

Démonstration. Tenant compte du fait que $\sum_{i=1}^{s} p_{i,\cdot} = \sum_{j=1}^{t} p_{\cdot,j} = 1$, il s'agit de tester que la loi sur E fait partie de la famille de lois caractérisées par (5.4) et paramétrées par les $s+t-2$ paramètres inconnus $p_{i,\cdot}$ $(i=1,\cdots,s-1)$ et $p_{\cdot,j}$ $(j=1,\cdots,t-1)$. On notera $\underline{p}$ le vecteur formé par ces paramètres. Cherchons l'expression de l'estimateur du maximum de vraisemblance du vecteur $\underline{p}$. On a

$$\sum_{(i,j)\in E} N_{i,j} = \sum_{i=1}^{s} N_{i,\cdot} = \sum_{j=1}^{t} N_{\cdot,j} = n$$

et, sous l'hypothèse H_0, la loi du vecteur $(N_{1,1}, \cdots, N_{s,t})$ est une loi multinômiale de paramètres $(n, (p_{1,\cdot}p_{\cdot,1}, \cdots, p_{s,\cdot}p_{\cdot,t}))$. La vraisemblance s'écrit donc

$$\begin{aligned} V_p(X_1, Y_1, \cdots, X_n, Y_n) &= P(N_{1,1} = n_{1,1}, \cdots, N_{s,t} = n_{s,t}) = n! \prod_{(i,j)\in E} \frac{(p_{i,\cdot}p_{\cdot,j})^{n_{i,j}}}{n_{i,j}!} \\ &= c \prod_{i=1}^{s} p_{i,\cdot}^{n_{i,\cdot}} \prod_{j=1}^{t} p_{\cdot,j}^{n_{\cdot,j}}, \end{aligned}$$

où c ne dépend pas des paramètres.
L'estimateur du maximum de vraisemblance maximise

$$L_p(X_1, X_n, \cdots, X_n, Y_n) = \ln c + \sum_{i=1}^{s-1} n_{i,\cdot} \ln p_{i,\cdot} + \sum_{j=1}^{t-1} n_{\cdot,j} \ln p_{\cdot,j} + n_{s,\cdot} \ln p_{s,\cdot} + n_{\cdot,t} \ln p_{\cdot,t}.$$

Soit en annulant les dérivées partielles de $L_p(X_1, \cdots, X_n)$,

$$\frac{n_{i,\cdot}}{p_{i,\cdot}} = \frac{n_{s,\cdot}}{p_{s,\cdot}} \quad \forall i = 1, \cdots, s-1 \quad \text{et} \quad \frac{n_{\cdot,j}}{p_{\cdot,j}} = \frac{n_{\cdot,t}}{p_{\cdot,t}} \quad \forall j = 1, \cdots, t-1,$$

ce qui équivaut à

$$\frac{n_{i,\cdot}}{p_{i,\cdot}} = \frac{n_{\cdot,j}}{p_{\cdot,j}} = n \quad \forall (i,j) \in E.$$

L'estimateur du maximum de vraisemblance du vecteur p est donc

$$\underline{\hat{p}}_{(n)} = {}^t\left(\frac{n_{1,\cdot}}{n}, \cdots, \frac{n_{s-1,\cdot}}{n}, \frac{n_{\cdot,1}}{n}, \cdots, \frac{n_{\cdot,t-1}}{n}\right).$$

Il suffit pour terminer d'appliquer la proposition 5.3. □

CHAPITRE 6

Les intervalles de confiance

Il est difficile d'imaginer qu'une statistique soit exactement égale à la valeur du paramètre qu'elle est censée estimer (si la loi de la statistique a une densité, cet événement est même de probabilité nulle). Donc, il est important que toute procédure d'estimation soit accompagnée d'une indication sur la précision de l'estimation.

1. Exemple

On considère un n-échantillon $X_1, \cdots, X_n$ de la loi gaussienne de paramètres (m, σ^2). On suppose que σ^2 est connu. On estime m. On a déjà vu toutes les qualités de l' estimateur $\overline{X}_n = n^{-1}\sum_{j=1}^n X_j$. On sait que la variable $\sqrt{n}\sigma^{-1}(\overline{X}_n - m)$ est gaussienne standard. Sa loi ne dépend pas du paramètre m, ce qu'on a déjà utilisé pour construire un test de niveau α pour tester $H_0 : m = m_0$ contre $m \neq m_0$ ou contre $m > m_0$. On sait par exemple que pour tout m

$$P_m\left(m - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \overline{X}_n \leq m + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha. \tag{6.1}$$

Ceci peut se lire aussi

$$P_m\left(\overline{X}_n - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq m \leq \overline{X}_n + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha, \tag{6.2}$$

ce qui change tout car dans (6.2), l'intervalle est aléatoire (la probabilité qu'il contienne le paramètre est $1-\alpha$), tandis que dans (6.1), l'intervalle est fixé et la variable aléatoire $\overline{X}_n$ a une probabilité $1-\alpha$ de se trouver dedans.

On dit que

$$I(X_1, \cdots, X_n) = [\overline{X}_n - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X}_n + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}]$$

est *un intervalle de confiance pour le paramètre* m. On dit que $1-\alpha$ est son niveau.

2. Remarques et liens avec les tests

REMARQUE 6.1. On a choisi un intervalle symétrique : est ce nécessaire ? Dans une certaine mesure la réponse est oui. En effet, soit U une variable gaussienne standard. On a vu que, pour $P(U \in [x, x+2z_{\alpha/2}])$ est maximale (et vaut $1-\alpha$) quand $x = -z_{\alpha/2}$. On en déduit que parmi les intervalles tels que $P(U \in [x, y]) = 1-\alpha$, le plus court est $[-z_{\alpha/2}, z_{\alpha/2}]$. Autrement dit l'intervalle proposé en (6.2) est le plus précis des intervalles de confiance de niveau $1-\alpha$ pour le paramètre m. Ceci dit, on peut avoir d'autres critères que la précision pour choisir l'intervalle de confiance. On peut vouloir par exemple une demi-droite de confiance si le seul souci est de garantir que le paramètre est suffisamment grand (ou suffisamment petit). |

REMARQUE 6.2. . Il y a à l'évidence un lien étroit entre test et intervalle de confiance. En effet, décider de rejeter l'hypothèse $H_0 : m = m_0$ lorsque $m_0 \notin I(X_1, \cdots, X_n)$ conduit exactement à la région de rejet que nous avions choisie pour tester cette hypothèse contre l'alternative $m \neq m_0$. |

3. Construction de régions de confiance

On va donner un principe général et quelques exemples simples.

3.1. Principe général. On cherche à estimer $\theta \in \Theta$ à partir de $(X_1, \cdots, X_n)$, variables indépendantes et de même loi P_θ. Ici, Θ est une partie borélienne de $\mathbb{R}^d$. Supposons qu'il existe une fonction $h(\theta, (x_1, \cdots, x_n))$ à valeurs dans $\mathbb{R}^k$ possédant les propriétés suivantes :
Propriété 1 : quelque soit θ, la fonction $h(\theta, (x_1, \cdots, x_n))$ est mesurable.
Propriété 2 : la loi de $h(\theta, (X_1, \cdots, X_n))$ ne dépend pas de θ.
Une telle fonction est parfois appelée *fonction pivotale*
Il est facile de s'appuyer sur une fonction pivotale pour construire des intervalles de confiance de la façon suivante. Supposons qu'il existe B, borélien de $\mathbb{R}^k$, tel que

$$P_\theta(h(\theta, (X_1, \cdots, X_n)) \in B) = 1 - \alpha \quad \forall \theta \in \Theta. \tag{6.3}$$

On prendra comme région de confiance

$$I(X_1, \cdots, X_n) = \{\theta \in \Theta | h(\theta, (X_1, \cdots, X_n)) \in B\}.$$

Comme dans l'exemple introductif, l'ensemble $I(X_1, \cdots, X_n)$ est aléatoire et a, sous P_θ, une probabilité $1 - \alpha$ de contenir θ. Dans l'exemple introductif on avait

$$h(m, (X_1, \cdots, X_n)) = \sqrt{n}\sigma^{-1}(\overline{X}_n - m),$$

et le borélien B était l'intervalle $[-z_{\alpha/2}, z_{\alpha/2}]$.

REMARQUE 6.3. L'égalité (3.6) s'écrit aussi $\pi(B) = 1-\alpha$ où π désigne la loi (ne dépendant pas de θ, d'après la Propriété 2) transportée de P_θ^n par $h(\theta, x)$. Bien sûr, un tel B peut ne pas exister. On recherchera alors un borélien tel que $\pi(B) > 1-\alpha$. |

REMARQUE 6.4. On peut aussi bien se servir de la fonction pivotale pour construire un test de l'hypothèse $\theta = \theta_0$. Pour cela il suffit de prendre comme région de rejet l'ensemble des $(X_1, \cdots, X_n)$ tels que $h(\theta_0, (X_1, \cdots, X_n)) \notin B$. Voir aussi la Remarque 6.2. |

3.2. Quelques exemples.

EXEMPLE 6.1. Estimation de l'espérance d'un échantillon gaussien de variance inconnue. Avec $s_n^2 = (n-1)^{-1} \sum_{j=1}^n (X_j - \overline{X}_n)^2$, on a vu que la loi de

$$h(m, (X_1, \cdots, X_n)) = \frac{\overline{X}_n - m}{s_n/\sqrt{n}}$$

ne dépend pas de m. C'est une loi de Student à $n-1$ degrés de liberté. On peut prendre $h(m, (x_1, \cdots, x_n))$ comme fonction pivotale, ce qui donne comme intervalle de confiance (symétrique) de niveau $1 - \alpha$

$$I(X_1, \cdots, X_n) = [\overline{X}_n - t_{n-1,\alpha/2}\frac{s_n}{\sqrt{n}}, \overline{X}_n + t_{n-1,\alpha/2}\frac{s_n}{\sqrt{n}}].$$

||

EXEMPLE 6.2. Estimation de la différence des espérances de deux échantillons gaussiens indépendants, de variances connues. On peut prendre comme fonction pivotale

$$h(m_1 - m_2, (X_1, \cdots, X_{n_1}, Y_1, \cdots, Y_{n_2})) = \frac{\overline{X}_{n_1} - \overline{Y}_{n_2} - m_1 + m_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

dont la loi est une gaussienne standard. Ceci conduit à un intervalle de confiance sur la différence des espérances :

$$I(X_1, \cdots, X_{n_1}, Y_1, \cdots, Y_{n_2}) = \Big[\overline{X}_{n_1} - \overline{Y}_{n_2} - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, \overline{X}_{n_1} - \overline{Y}_{n_2} + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\Big]$$

‖

EXEMPLE 6.3. Estimation de la différence des espérances de deux échantillons gaussiens indépendants, de variances égales mais inconnues.
La fonction

$$h(m_1 - m_2, (X_1, \cdots, X_{n_1}, Y_1, \cdots, Y_{n_2})) = \frac{\overline{X}_{n_1} - \overline{Y}_{n_2} - m_1 + m_2}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\sqrt{\frac{\sum_{j=1}^{n_1}(X_j - \overline{X}_{n_1})^2 + \sum_{j=1}^{n_2}(Y_j - \overline{Y}_{n_2})^2}{n_1 + n_2 - 2}}}$$

est distribuée selon une loi de Student à $n_1 + n_2 - 2$ degrés de liberté, et donne à nouveau un intervalle de confiance sur $m_1 - m_2$. Le construire.
‖

EXEMPLE 6.4. Estimation de la différence des espérances de deux échantillons gaussiens appariés.
Avec $D_j = X_j - Y_j$ et $s_n^2 = (n-1)^{-1}\sum_{j=1}^n (D_j - \overline{D}_n)^2$, prendre comme fonction pivotale

$$h(m_1 - m_2, (X_1, Y_1) \cdots, (X_n, Y_n)) = \frac{\overline{D}_n - m}{s_n/\sqrt{n}},$$

dont la loi est une loi de Student à $n-1$ degrés de liberté. Construire l'intervalle de confiance. ‖

EXEMPLE 6.5. Estimation de la variance d'un échantillon gaussien d'espérance inconnue. On peut prendre comme fonction pivotale

$$h(\sigma^2, (X_1, \cdots, X_n)) = \frac{\sum_{j=1}^n (X_j - \overline{X}_n)^2}{\sigma^2},$$

qui est distribuée selon une loi du χ^2 à $n-1$ degrés de liberté. Soit un intervalle $[a_\alpha, b_\alpha]$ tel que $P(\chi^2_{n-1} \in [a_\alpha, b_\alpha]) = 1 - \alpha$. L'intervalle

$$I(X_1, \cdots, X_n) = \Big[\frac{\sum_{j=1}^n (X_j - \overline{X}_n)^2}{b_\alpha}, \frac{\sum_{j=1}^n (X_j - \overline{X}_n)^2}{a_\alpha}\Big]$$

est clairement un intervalle de confiance de niveau $1 - \alpha$ pour le paramètre σ^2. ‖

EXEMPLE 6.6. Estimation du couple (espérance, variance) pour un échantillon gaussien.
Considérons la fonction à valeurs dans $\mathbb{R}^2$

$$h(\sigma^2, m, (X_1, \cdots, X_n)) = \begin{pmatrix} \sqrt{n}\frac{\overline{X}_n - m}{\sigma} \\ \frac{\sum_{j=1}^n (X_j - \overline{X}_n)^2}{\sigma^2} \end{pmatrix}$$

Sa loi ne dépend pas des paramètres car sa première coordonnée a une loi gaussienne standard, sa seconde coordonnée est un χ^2_{n-1}, et ces coordonnées sont indépendantes. Soient α_1 et α_2 tels que $(1-\alpha_1)(1-\alpha_2) = 1-\alpha$. Considérons les deux bornes a_{α_2} et b_{α_2} de la question précédente. On a

$$P_{m,\sigma^2}\left(-z_{\alpha_1/2} \leq \sqrt{n}\frac{\overline{X}_n - m}{\sigma} \leq z_{\alpha_1/2}\ ,\ a_{\alpha_2} \leq \frac{\sum_{j=1}^n (X_j - \overline{X}_n)^2}{\sigma^2} \leq b_{\alpha_2}\right)$$
$$= \quad (1-\alpha_1)(1-\alpha_2) = 1-\alpha.$$

Donc, le domaine

$$\left\{\frac{\sum_{j=1}^n (X_j - \overline{X}_n)^2}{b_{\alpha_2}} \leq \sigma^2 \leq \frac{\sum_{j=1}^n (X_j - \overline{X}_n)^2}{b_{\alpha_2}}\ ,\ n\frac{(\overline{X}_n - m)^2}{\sigma^2} \leq z^2_{\alpha_1/2}\right\}$$

est un domaine de confiance de niveau $1-\alpha$ pour le couple (m, σ^2). Le tracer (il est délimité par deux droites et une parabole). ‖

CHAPITRE 7

Régression linéaire

Il s'agit ici de variables indépendantes mais non identiquement distribuées. Leurs distributions diffèrent seulement par leurs espérances, qui dépendent *linéairement* de paramètres inconnus.

1. Introduction : les modèles

1.1. Le modèle de régression simple. C'est, comme son nom l'indique, le plus simple :

$$y_j = a + bx_j + \varepsilon_j \qquad j = 1, \cdots, n. \tag{7.1}$$

Dans cette expression, les y_j représentent par exemple la consommation annuelle de n individus. Ce sont des variables aléatoires *observées*.
Les x_j sont parfois appelées *variables explicatives* ou encore *régresseurs*. Elles représentent par exemple le revenu annuel des individus étudiés . Ce sont des quantités connues et *non aléatoires* à qui on demande de n'être pas toutes égales.
Les ε_j sont des variables aléatoires *non observées*. On dit parfois que $(\varepsilon_j)_{j\geq 1}$ est un bruit.
Les paramètres a et b sont inconnus et il s'agit de les estimer. C'est pour cela qu'on suppose que les x_j ne sont pas toutes égales. En effet, dans le cas contraire les équations (7.1) s'écriraient $y_j = a + bx + \varepsilon_j \quad (j = 1, \cdots, n)$, et on voit mal comment on pourrait parvenir à estimer autre chose que $a + bx$.
Plus loin, on supposera que les variables ε_j sont indépendantes, toutes de même loi et centrées et de variance inconnue. Il faudra donc estimer leur variance.

1.2. La régression multiple. On complique le modèle en introduisant plusieurs régresseurs (par exemple, au revenu on adjoint l'âge)

$$y_j = \sum_{l=1}^{k} b_l x_{j,l} + \varepsilon_j \qquad j = 1, \cdots, n. \tag{7.2}$$

Le nombre de régresseurs k est bien entendu inférieur au nombre d'observations n.
Le modèle de régression simple est un cas particulier obtenu pour $k = 2$ et $x_{j,1} = 1 \quad \forall j$. Les hypothèses sur le bruit sont inchangées.
Il est plus commode d'écrire ce modèle sous la forme vectorielle

$$Y = \mathbf{X}B + \varepsilon = b_1 X_1 + \cdots + b_k X_k + \varepsilon, \tag{7.3}$$

où X_j désigne le vecteur ${}^t(x_{1,j}, \cdots, x_{n,j})$ et où

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \quad \mathbf{X} = \begin{pmatrix} x_{1,1} & \dots & x_{1,k} \\ \vdots & \vdots & \vdots \\ x_{n,1} & \dots & x_{n,k} \end{pmatrix} \quad B = \begin{pmatrix} b_1 \\ \vdots \\ b_k \end{pmatrix} \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

En particulier, le modèle de régression simple s'écrit de la même façon avec

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}.$$

La contrainte mise sur les régresseurs x_j au paragraphe précédent signifie simplement que la matrice $\mathbf{X}$ est de rang 2. Dans le cas général (7.3) on supposera que cette matrice est de rang k. On verra plus loin l'intérêt de cette hypothèse, mais on peut déjà comprendre que dans le cas contraire la décomposition $b_1X_1 + \cdots + b_kX_k$ n'est pas unique, et donc le modèle n'est pas pertinent.

2. Estimation des moindres carrés

2.1. Le principe et les résultats. Une très ancienne technique pour estimer les paramètres b_j, dite *méthode des moindres carrés*, consiste à minimiser, par rapport aux variables z_j, la somme des carrés

$$\sum_{j=1}^{n}(y_j - z_1x_{j,1} - \cdots - z_kx_{j,k})^2 = \|Y - \mathbf{X}Z\|_2^2 = \|Y - z_1X_1 - \cdots - z_kX_k\|_2^2.$$

Cette méthode consiste donc simplement à projeter orthogonalement Y sur $[X_1, \cdots, X_k]$, espace engendré dans $\mathbb{R}^n$ par les k vecteurs X_j que l'on a supposés linéairement indépendants.
Le résultat s'obtient en annulant les dérivées partielles de la forme quadratique

$$\sum_{j=1}^{n}(y_j - z_1x_{j,1} - \cdots - z_kx_{j,k})^2.$$

En effet cette forme quadratique, positive et indéfiniment dérivable, atteint un minimum en un point qui annule ses dérivées partielles. Cette annulation conduit à la résolution du système de k équations linéaires

$$\sum_{j=1}^{n} x_{j,l}(y_j - z_1x_{j,1} - \cdots - z_kx_{j,k}) = 0 \quad \forall l = 1, \cdots, k,$$

soit aussi

$${}^t\mathbf{X}\mathbf{X}Z = {}^t\mathbf{X}Y.$$

L'estimateur de B s'écrit donc

$$\hat{B}_n = ({}^t\mathbf{X}\mathbf{X})^{-1}\ {}^t\mathbf{X}Y. \tag{7.4}$$

Remarque 7.1. Au passage, on remarque que le projeté othogonal de Y sur $[X_1, \cdots, X_k]$ s'écrit

$$\mathbf{X}\hat{B}_n = \mathbf{X}({}^t\mathbf{X}\mathbf{X})^{-1}\ {}^t\mathbf{X}Y. \tag{7.5}$$

De ce fait, la matrice de l'opérateur de projection orthogonale dans $\mathbb{R}^n$ sur l'espace $[X_1, \cdots, X_k]$ est

$$M := \mathbf{X}\ ({}^t\mathbf{X}\mathbf{X})^{-1}\ {}^t\mathbf{X}. \tag{7.6}$$

On note que, si I désigne la matrice de l'opérateur identité de $\mathbb{R}^n$, $I - M$ est la matrice de la projection orthogonale sur le supplémentaire orthogonal de $[X_1, \cdots, X_k]$. De ce fait

$$(I - M)\mathbf{X} = 0, \tag{7.7}$$

résultat qui sera utile par la suite.

Dans le cas particulier de la régression simple on a

$$ {}^t\mathbf{X}\mathbf{X} = \begin{pmatrix} n & \sum_{j=1}^n x_j \\ \sum_{j=1}^n x_j & \sum_{j=1}^n x_j^2, \end{pmatrix}, $$

d'où résulte en notant $\overline{x}_n$ et $\overline{y}_n$ les moyennes arithmétiques respectives des x_j et des y_j :

$$ \hat{a}_n = \frac{\overline{y}_n \sum_{j=1}^n x_j^2 - \overline{x}_n \sum_{j=1}^n x_j y_j}{n\sum_{j=1}^n x_j^2 - n\overline{x}_n^2} \tag{7.8} $$

$$ \hat{b}_n = \frac{\sum_{j=1}^n (x_j - \overline{x}_n)(y_j - \overline{y}_n)}{\sum_{j=1}^n (x_j - \overline{x}_n)^2}. \tag{7.9} $$

La droite d'équation $y = \hat{a}_n + \hat{b}_n x$ s'appelle droite de régression du vecteur Y sur le vecteur ${}^tX = {}^t(x_1, \cdots, x_n)$, sa pente $\hat{b}_n$ s'appelle coefficient de régression et on vérifie que

$$ \hat{a}_n = \overline{y}_n - \hat{b}_n \overline{x}_n, $$

ce qui signifie que la droite de régression passe par le point moyen $(\overline{x}_n, \overline{y}_n)$.

2.2. Résidu, estimation de la variance et moments des estimateurs. On note

$$ R_n = Y - \mathbf{X}\hat{B}_n $$

la différence entre Y et son projeté. En fait, R_n est le projeté orthogonal de Y sur le supplémentaire orthogonal de $[X_1, \cdots, X_k]$ dans $\mathbb{R}^n$.

PROPOSITION 7.1. *Si les variables ε_j sont centrées et mutuellement indépendantes,*

- *L'estimateur des moindres carrés $\hat{B}_n$ est un estimateur sans biais de B. Sa matrice de covariance est*

$$ \Sigma_{\hat{B}_n} = \sigma^2 ({}^t\mathbf{X}\mathbf{X})^{-1}, \tag{7.10} $$

- *le résidu R_n est centré, sa matrice de covariance est*

$$ \Sigma_{R_n} = \sigma^2 (I - M). \tag{7.11} $$

- *$\hat{B}_n$ et R_n sont orthogonaux dans l'espace des vecteurs aléatroires de L^2, c'est à dire*

$$ \mathbb{E}(\hat{B}_n \, {}^t R_n) = 0 $$

DÉMONSTRATION. On déduit de (7.3) que $\mathbb{E}Y = \mathbf{X}B$. En conséquence (7.4) conduit à

$$ \mathbb{E}\hat{B}_n = ({}^t\mathbf{X}\mathbf{X})^{-1} \, {}^t\mathbf{X}\mathbb{E}Y = B. $$

Puis,

$$ \begin{aligned} \mathbb{E}\left((\hat{B}_n - B)^t(\hat{B}_n - B)\right) &= \mathbb{E}\left(({}^t\mathbf{X}\mathbf{X})^{-1} \, {}^t\mathbf{X}(Y - \mathbf{X})^t(Y - \mathbf{X})\mathbf{X}({}^t\mathbf{X}\mathbf{X})^{-1}\right) \\ &= ({}^t\mathbf{X}\mathbf{X})^{-1} \, {}^t\mathbf{X}\Sigma_Y\mathbf{X}({}^t\mathbf{X}\mathbf{X})^{-1}. \end{aligned} $$

Comme $\Sigma_Y = \Sigma_\varepsilon = \sigma^2 I$, (7.10) est prouvée.
Considérons maintenant le résidu. On a

$$ \mathbb{E}R_n = \mathbb{E}(Y - \mathbf{X}\hat{B}_n) = \mathbb{E}Y - \mathbf{X}\mathbb{E}\hat{B}_n = 0. $$

Pour les moments d'ordre deux :

$$\Sigma_{R_n} = \mathbb{E}\left(R_n\ {}^tR_n\right) = \mathbb{E}(I-M)Y^tY(I-M) = (I-M)\mathbb{E}(Y^tY)(I-M).$$

Mais $\mathbb{E}(Y^tY) = \sigma^2 I + \mathbb{E}Y\mathbb{E}^tY = \sigma^2 I + \mathbf{X}B^tB^t\mathbf{X}$. Comme $(I-M)X = 0$, on obtient bien (7.11). Enfin,

$$\begin{aligned}\mathbb{E}(\hat{B}_n\ {}^tR_n) &= \mathbb{E}\left(\hat{B}_n\ {}^t(Y-\mathbf{X}\hat{B}_n)\right) = \mathbb{E}\left(\hat{B}_n\ {}^tY(I-M)\right)\\ &= \mathbb{E}\left({}^t\mathbf{X}\mathbf{X})^{-1}\ {}^t\mathbf{X}Y^tY(I-M)\right) = \sigma^2\left({}^t\mathbf{X}\mathbf{X})^{-1}\ {}^t\mathbf{X}(I-M)\right) = 0.\end{aligned}$$

La proposition est donc prouvée. □

Examinons maintenant la norme du résidu :

$$\begin{aligned}\|R_n\|_2^2 =& \sum_{j=1}^{n}(y_j - \hat{b}_1 x_{j,1} - \cdots - \hat{b}_j x_{j,k})^2 = \|Y - \mathbf{X}\hat{B}_n\|_2^2\\ =& \|Y - \hat{b}_1 X_1 - \cdots - \hat{b}_k X_k\|_2^2.\end{aligned}$$

La proposition suivante montre l'on peut construire un estimateur sans biais de la variance σ^2 à partir de $\|R_n\|_2^2$.

PROPOSITION 7.2. *Si les variables ε_j sont centrées, mutuellement indépendantes et de variance σ^2, on a*

$$\mathbb{E}\|R_n\|_2^2 = (n-k)\sigma^2.$$

DÉMONSTRATION. en utilisant (7.5) et (7.6), on a

$$\begin{aligned}\|R_n\|_2^2 &= {}^t(Y-\mathbf{X}\hat{B}_n)(Y-\mathbf{X}\hat{B}_n) = {}^t\,(Y-MY)(Y-MY)\\ &= {}^tY(I-M)^2Y = tr\left((I-M)Y^tY(I-M)\right),\end{aligned}$$

où la dernière inégalité vient du fait que pour tout vecteur v on a

$${}^tvv = \sum v_j^2 = tr(v^tv).$$

Puis, parce que la trace de l'espérance est égale à l'espérance de la trace,

$$\begin{aligned}\mathbb{E}\|R_n\|_2^2 &= \mathbb{E}tr\left((I-M)Y^tY(I-M)\right) = tr\left(\mathbb{E}\left((I-M)Y^tY(I-M)\right)\right)\\ &= \sigma^2 tr\left((I-M)^2\right) = \sigma^2 tr(I-M) = (n-k)\sigma^2,\end{aligned}$$

ce qui termine la preuve. □

Retour à la régression simple.
Dans ce cas $\|R_n\|_2^2$ a une expression explicite. C'est en effet le carré de la norme du vecteur $Y - \mathbf{X}\hat{B}_n = Y - MY$. On prouve que

$$\|R_n\|_2^2 = s_{X,Y}^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

où $s_{x,y}$ est le sinus de l'angle que font les deux vecteurs X et Y recentrés à leurs moyennes. Soit

$$s_{x,y}^2 = 1 - r_{x,y}^2 = 1 - \frac{\left(\sum_{j=1}^{n}(x_j - \overline{x}_n)(y_j - \overline{y}_n)\right)^2}{\sum_{j=1}^{n}(x_j - \overline{x}_n)^2 \sum_{j=1}^{n}(y_j - \overline{y}_n)^2}.$$

Le coefficient $r_{x,y}$, cosinus de l'angle est appelé cœfficient de corrélation entre les deux vecteurs. Son égalité à ± 1 signifie que les deux vecteurs sont co-linéaires. Son égalité à zéro qu'ils sont orthogonaux.

2.3. Théorème de Gauss Markov. Soit v un vecteur fixé de $\mathbb{R}^k$.
On va regarder d'abord ${}^tv\hat{B}_n$. C'est un estimateur sans biais du paramètre réel ${}^tvB$. On sait que ${}^tv\hat{B}_n = ({}^tv{}^t\mathbf{X}\mathbf{X})^{-1}\,{}^t\mathbf{X}Y = {}^tw_0Y$. Considérons l'ensemble $\mathcal{E}_v$ des estimateurs sans biais de ${}^tvB$ qui sont, comme ${}^tv\hat{B}_n$, des transformés linéaires de Y *i.e.*

$$\mathcal{E}_v = \left\{ {}^twY \ : w \in \mathbb{R}^k \text{ et } \mathbb{E}({}^twY) = {}^t\,vB \right\}$$

THÉORÈME 7.3. *Théorème de Gauss-Markov*
${}^tv\hat{B}_n$ *est l'unique élément de* $\mathcal{E}_v$ *de variance minimum.*

DÉMONSTRATION. soit ${}^twY$ un élément de $\mathcal{E}_v$. On a,

$$\mathbb{E}\left(({}^tw_0 - {}^tw)Y\right) = 0 = \left({}^tw_0 - {}^tw\right)\mathbf{X}B \qquad \forall B,$$

ce qui équivaut à

$$\left({}^tw_0 - {}^tw\right)\mathbf{X} = 0.$$

Calculons la variance de cet estimateur.

$$\begin{aligned}\mathrm{Var}({}^twY) &= \mathrm{Var}\left({}^tw_0Y + ({}^tw - {}^tw_0)Y\right)\\ &= \mathrm{Var}({}^tw_0Y) + \mathrm{Var}({}^t(w-w_0)Y) + 2\mathrm{Cov}({}^tw_0Y,\ {}^t(w-w_0)Y).\end{aligned}$$

Mais

$${}^t(w-w_0)Y = {}^t(w-w_0)(\mathbf{X}B+\varepsilon) = {}^t(w-w_0)\varepsilon,$$

si bien que

$$\begin{aligned}\mathrm{Cov}({}^tw_0Y,\ {}^t(w-w_0)Y) &= \mathrm{Cov}({}^tw_0Y,\ {}^t(w-w_0)\varepsilon) = \mathbb{E}\left({}^tw_0Y\ {}^t\varepsilon(w-w_0)\right)\\ &= \mathbb{E}\left({}^tw_0\varepsilon\ {}^t\varepsilon(w-w_0)\right) = \mathbb{E}\left({}^tw_0(w-w_0)\right)\\ &= \mathbb{E}\left({}^tv({}^t\mathbf{X}\mathbf{X})^{-1}\,{}^t\mathbf{X}(w-w_0)\right) = 0.\end{aligned}$$

On en déduit que

$$\mathrm{Var}({}^twY) \geq \mathrm{Var}({}^tw_0Y),$$

l'égalité ayant lieu si et seulement si $\mathrm{Var}({}^t(w-w_0)Y) = 0$, c'est à dire si et seulement si

$$\mathrm{Var}({}^t(w-w_0)\varepsilon) = \sigma^2\,{}^t(w-w_0)(w-w_0) = 0,$$

ce qui signifie que les deux vecteurs w et w_0 sont égaux. □

On peut obtenir facilement une version multidimensionnelle de ce théorème. Soit U une matrice $r \times k$. Considérons le vecteur UB. Il est clair que

$$U\hat{B}_n = U({}^t\mathbf{X}\mathbf{X})^{-1}\,{}^t\mathbf{X}Y = L_0Y$$

est un estimateur sans biais de UB et qu'il est obtenu par transformation linéaire de Y. Considérons $\mathcal{E}_U$ l'ensemble des estimateurs qui ont les mêmes propriétés.

COROLLAIRE 7.4. *Pour tout élément* LY *de* $\mathcal{E}_V$*, sa matrice de covariance est supérieure ou égale à celle de* $U\hat{B}_n$*, ce qui signifie que la matrice symétrique* $\Sigma_{LY} - \Sigma_{L_0Y}$ *est semi-définie positive.*

DÉMONSTRATION. Il faut montrer que, pour tout vecteur w de $\mathbb{R}^r$,

$${}^tw\left(\Sigma_{LY} - \Sigma_{L_0Y}\right)w \geq 0.$$

or, ${}^tw\left(\Sigma_{LY}\right)w$ et ${}^tw\left(\Sigma_{L_0Y}\right)w$ sont les variances respectives de ${}^twLY$ et de ${}^twL_0Y = {}^twU\hat{B}_n$. Le théorème de Gauss-Markov, permet de conclure. □

En particulier, en prenant U la matrice identité, on obtient

COROLLAIRE 7.5. *$\hat{B}_n$ est de matrice de covariance minimale parmi les estimateurs de B qui sont transformés linéaires de Y.*

3. Convergence des estimateurs

On va ici s'occuper de l'estimateur de B.

3.1. Convergence en moyenne quadratique.

PROPOSITION 7.6. *L'estimateur des moindres carrés $\hat{B}_n$ est convergent en moyenne quadratique si et seulement si la plus petite valeur propre de la matrice ${}^t\mathbf{XX}$ tend vers l'infini quand $n \to \infty$.*

DÉMONSTRATION. Comme $\hat{B}_n$ est non biaisé, sa convergence en moyenne quadratique s'écrit

$$\mathbb{E}\|\hat{B}_n - B\|_2^2 \to 0.$$

Comme $\|\hat{B}_n - B\|_2^2 = tr\left((\hat{B}_n - B)\ {}^t(\hat{B}_n - B)\right)$, cette convergence a lieu si et seulement si

$$tr\left(\Sigma_{\hat{B}_n}\right) = \sigma^2 tr({}^t\mathbf{XX})^{-1} \to 0,$$

autrement dit si et seulement si la plus grande valeur propre de la matrice $({}^t\mathbf{XX})^{-1}$ tend vers zéro (ou bien de façon équivalente si la plus petite valeur propre de ${}^t\mathbf{XX}$ tend vers l'infini) quand k tend vers l'infini. □

Notamment, dans le cas de la régression simple, on a

$${}^t\mathbf{XX} = \begin{pmatrix} n & \sum_{j=1}^n x_j \\ \sum_{j=1}^n x_j & \sum_{j=1}^n x_j^2, \end{pmatrix}.$$

On vérifiera que la plus petite valeur propre tend vers l'infini si et seulement si $n\sum_{j=1}^n (x_j - \overline{x}_n)^2 \to \infty$.

3.2. Convergence en loi. On ne considère ici que la convergence de l'estimateur de la pente de la droite des moindres carrés dans le cas de la régression simple.

PROPOSITION 7.7. *Supposons que les ε_j sont indépendants, centrés, de variance σ^2 finie et tous de même loi. Si*

$$\gamma_n^2 = \max\left\{\frac{(x_j - \overline{x}_n)^2}{\sum_{j=1}^n (x_j - \overline{x}_n)^2} | j = 1, \cdots, n\right\} \to 0 \quad \textit{quand} \quad n \to \infty, \tag{7.12}$$

alors, $\hat{b}_n$ étant défini en (7.8) la suite

$$\sqrt{\frac{\sum_{j=1}^n (x_j - \overline{x}_n)^2}{\sigma^2}}(\hat{b}_n - b)$$

converge en loi vers une variable gaussienne standard.

Preuve. Il est facile de vérifier à partir de (7.8) et en revenant aux équations (7.1) que

$$\begin{aligned}\hat{b}_n &= \frac{\sum_{j=1}^n (x_j-\overline{x}_n)(y_j-\overline{y}_n)}{\sum_{j=1}^n (x_j-\overline{x}_n)^2} = \frac{\sum_{j=1}^n (x_j-\overline{x}_n)y_j}{\sum_{j=1}^n (x_j-\overline{x}_n)^2} \\ &= \frac{b\sum_{j=1}^n (x_j-\overline{x}_n)^2 + \sum_{j=1}^n (y_j-\overline{y}_n)\varepsilon_j}{\sum_{j=1}^n (x_j-\overline{x}_n)^2} \\ &= b + \frac{\sum_{j=1}^n (x_j-\overline{x}_n)\varepsilon_j}{\sum_{j=1}^n (x_j-\overline{x}_n)^2}.\end{aligned}$$

Montrons que les conditions du théorème de Lindberg-Feller sont vérifiées. Ce théorème s'énonce de la façon suivante :

THÉORÈME 7.8 (Théorème de Lindberg-Feller). *Soit un tableau triangulaire* $(X_{n,1},\cdots,X_{n,k_n})_{n\in\mathbb{N}}$ *de variables telles que*

- *pour* n *fixé les variables* $(X_{n,1},\cdots,X_{n,k_n})$ *sont indépendantes*
- *toutes les variables ont un second moment fini.*
- *notant* $D_n^2 = \sum_{j=1}^{k_n} \mathrm{Var}X_{n,j}$, *on a pour tout* $\tau > 0$

$$\lim_{n\to\infty} \frac{1}{D_n^2} \sum_{j=1}^{k_n} \mathbb{E}\left((X_{n,j}-\mathbb{E}(X_{n,j}))^2 \mathbb{I}_{]\tau D_n,\infty[}(|X_{n,j}-\mathbb{E}(X_{n,j})|)\right) = 0.$$

Alors, la suite

$$\frac{1}{D_n}\sum_{j=1}^{k_n}(X_{n,j}-\mathbb{E}(X_{n,j}))$$

converge en loi vers une variable gaussienne standard.

DÉMONSTRATION. *(Suite de preuve de la proposition 7.7.)*

Appliquons le théorème de Lindberg-Feller à la suite $X_{n,j} = (x_j-\overline{x}_n)\varepsilon_j$, avec $k_n = n$. On a

$$\mathrm{Var}X_{n,j} = \mathrm{Var}\left((x_j-\overline{x}_n)\varepsilon_j\right) = \sigma^2(x_j-\overline{x}_n)^2$$

d'où

$$D_n^2 = \sigma^2\sum_{j=1}^n (x_j-\overline{x}_n)^2.$$

De plus,

$$\begin{aligned}&\mathbb{E}\left((X_{n,j}-\mathbb{E}(X_{n,j}))^2\mathbb{I}_{]\tau D_n,\infty[}(|X_{n,j}-\mathbb{E}(X_{n,j})|)\right) \\ =\ &(x_j-\overline{x}_n)^2\mathbb{E}\left(\varepsilon_j^2\mathbb{I}_{]\tau D_n,\infty[}(|\varepsilon_j(x_j-\overline{x}_n)|)\right) \\ =\ &(x_j-\overline{x}_n)^2\mathbb{E}\left(\varepsilon_1^2\mathbb{I}_{]\tau D_n,\infty[}(|\varepsilon_1(x_j-\overline{x}_n)|)\right).\end{aligned}$$

Puis, remarquant que

$$|\varepsilon_1(x_j-\overline{x}_n)| > \tau D_n \Longrightarrow |\varepsilon_1| > \frac{\tau\sigma}{\gamma_n},$$

on obtient

$$\mathbb{E}\left((X_{n,j}-\mathbb{E}(X_{n,j}))^2\mathbb{I}_{]\tau D_n,\infty[}(|X_{n,j}-\mathbb{E}(X_{n,j})|)\right) \le (x_j-\overline{x}_n)^2\mathbb{E}\left(\varepsilon_1^2\mathbb{I}_{]\frac{\tau\sigma}{\gamma_n},\infty[}(|\varepsilon_1|)\right).$$

D'où finalement

$$\begin{aligned}&\frac{1}{D_n^2}\sum_{j=1}^n \mathbb{E}\left((X_{n,j}-\mathbb{E}(X_{n,j}))^2\mathbb{I}_{]\tau D_n,\infty[}(|X_{n,j}-\mathbb{E}(X_{n,j})|)\right)\\ \leq\ &\frac{1}{D_n^2}\sum_{j=1}^n (x_j-\overline{x}_n)^2\mathbb{E}\left(\varepsilon_1^2\mathbb{I}_{]\frac{\tau\sigma}{\gamma_n},\infty[}(|\varepsilon_1|)\right)=\frac{1}{\sigma^2}\mathbb{E}\left(\varepsilon_1^2\mathbb{I}_{]\frac{\tau\sigma}{\gamma_n},\infty[}(|\varepsilon_1|)\right).\end{aligned}$$

L'hypothèse (7.12) entraîne que $\frac{\tau}{\sqrt{\gamma}_n}\to\infty$. Comme $\mathbb{E}(\varepsilon_1^2)<\infty$, la convergence vers zéro s'ensuit. □

4. Le cas gaussien

On va maintenant supposer que les ε_j sont indépendantes et toutes de même loi gaussienne centrée de variance σ^2. Les variables y_j sont donc elles aussi indépendantes, gaussiennes. Elles ont même variance σ^2 et $\mathbb{E}y_j=\sum_{l=1}^k b_l x_{j,l}$. En conséquence la densité de $(y_1,\cdots,y_n)$ est

$$g(y_1,\cdots,y_n) = \frac{1}{(2\pi\sigma^2)^{n/2}}\exp-\frac{\sum_{j=1}^n(y_j-b_1x_{j,1}-\cdots-b_kx_{j,k})^2}{2\sigma^2} \quad (7.13)$$

$$= \frac{1}{(2\pi\sigma^2)^{n/2}}\exp-\frac{\|Y-\mathbf{X}B\|_2^2}{2\sigma^2} \quad (7.14)$$

4.1. Estimateur du maximum de vraisemblance. Cet estimateur minimise

$$\frac{\|Y-\mathbf{X}B\|_2^2}{2\sigma^2}-\frac{n}{2}\ln(2\pi\sigma^2)$$

On sait que $\|Y-\mathbf{X}B\|_2^2$ est minimum si B prend la valeur $\hat{B}_n$. Donc on trouve ici que l'estimateur des moindres carrés est celui du maximum de vraisemblance. Puis, en annulant la dérivée par rapport à σ^2, on obtient

$$\hat{\sigma^2}_n^{mv}=\frac{\|Y-\mathbf{X}\hat{B}_n\|_2^2}{n},$$

qui est biaisé, mais asymptotiquement sans biais, comme on l'a vu dans la proposition 7.2.

4.2. Exhaustivité.

PROPOSITION 7.9. *La statistique $(\hat{B}_n,\hat{\sigma^2}_n)$ est exhaustive et totale.*

DÉMONSTRATION. on ne prouvera que l'exhaustivité. Puisque $Y-\mathbf{X}\hat{B}_n$ est orthogonal aux vecteurs $X_1,\cdots,X_k$, on a

$$\|Y-\mathbf{X}B\|_2^2=\|Y-\mathbf{X}\hat{B}_n\|_2^2+\|\mathbf{X}B-\mathbf{X}\hat{B}_n\|_2^2.$$

La vraisemblance s'écrit donc

$$g(y_1,\cdots,y_n)=\frac{1}{(2\pi\sigma^2)^{n/2}}\exp-\frac{(n-k)\hat{\sigma^2}_n+\|\mathbf{X}B-\mathbf{X}\hat{B}_n\|_2^2}{2\sigma^2},$$

ce qui prouve l'exhaustivité. □

Une conséquence de cette proposition est que le couple $(\hat{B}_n,\hat{\sigma^2}_n)$ est optimal dans la classe des estimateurs sans biais. Ce résultat est à comparer, en ce qui concerne $\hat{B}_n$, avec le théorème de Gauss-Markov qui donne seulement l'optimalité de $\hat{B}_n$ dans la classe des estimateurs sans biais *transformés linéaires de Y*.

4.3. Lois des estimateurs. On a vu au début de la section que les variables y_j sont indépendantes et gaussiennes. Le vecteur Y est donc aussi gaussien. Sa matrice de covariance est $\sigma^2 I_n$ et le vecteur des espérances est $\mathbf{X}B$. On remarque que le vecteur ${}^t({}^t\hat{B}_n, {}^tR_n)$ s'écrit

$$\begin{pmatrix} \hat{B}_n \\ R_n \end{pmatrix} = \begin{pmatrix} ({}^t\mathbf{X}\mathbf{X})^{-1}\,{}^t\mathbf{X} \\ I - M \end{pmatrix} Y.$$

Donc ce vecteur est aussi gaussien, son espérance est ${}^t({}^tB, 0)$ et sa matrice de covariance, d'après la proposition 23, est

$$\begin{pmatrix} \sigma^2({}^t\mathbf{X}\mathbf{X})^{-1} & 0 \\ 0 & \sigma^2(I-M) \end{pmatrix}.$$

De plus, on se rappelle que R_n est le projeté orthogonal de Y sur le supplémentaire orthogonal de $[X_1, \cdots, X_k]$ dans $\mathbb{R}^n$.
De ceci on déduit les résultats suivants :

PROPOSITION 7.10. *Si les (ε_j) sont gaussiens, indépendants, centrés et tous de variance σ^2,*

- *$\hat{B}_n$ est gaussien, d'espérance B et de matrice de covariance $\sigma^2({}^t\mathbf{X}\mathbf{X})^{-1}$,*
- *$\hat{B}_n$ et $\hat{\sigma^2}_n$ sont indépendants,*
- *$\sigma^{-2}(n-k)\hat{\sigma^2}_n$ est distribué selon une loi de χ^2_{n-k}.*

L'intérêt de ce résultat est qu'il permet la constructions de tests et de régions de confiance pour les paramètres.