

# Statistique et modèles aléatoires

Dominique Picard

16 mars 2004



# Chapitre 1

## 1. METHODOLOGIE STATISTIQUE

### 1.1 Introduction. Modélisation statistique.

À la base, le statisticien dispose d'une **observation**  $x$  à valeurs dans un espace  $\mathcal{X}$ . La modélisation consiste à faire l'hypothèse que cette observation est la réalisation  $X(\omega)$  d'une variable aléatoire  $X$ , à valeurs dans  $(\mathcal{X}, \mathcal{A})$ . ( $\mathcal{A}$  est alors une tribu sur  $\mathcal{X}$  et  $\omega$  appartient à un ensemble  $\Omega$ .)

Formellement, on a un triplet  $(\Omega, \mathcal{F}, P)$  c'est à dire que  $\mathcal{F}$  est une tribu sur  $\Omega$ , et  $P$  une mesure de probabilité sur  $\mathcal{F}$ .

$X$  est une application mesurable de  $(\Omega, \mathcal{F})$  dans  $(\mathcal{X}, \mathcal{A})$ , et la loi de  $X$ ,  $P_X$ , est la mesure image de  $P$  par  $X$  définie pour tout ensemble  $A$  de  $\mathcal{A}$  par la formule

$$P^X(A) = P(X^{-1}(A)).$$

En statistique, nous verrons que  $\mathcal{X}$  est souvent  $\mathbb{R}^n$  ou un sous ensemble de  $\mathbb{R}^n$ ,  $\mathcal{A}$  est généralement sa tribu borelienne, et l'espace  $(\Omega, \mathcal{F})$  joue un rôle très auxiliaire. La plupart du temps, on peut identifier  $(\Omega, \mathcal{F})$  et  $(\mathcal{X}, \mathcal{A})$ , de sorte que  $X$  devient l'application identité. On dira alors que le modèle est en position canonique.

**Définition 1** On appelle **Modèle statistique** ou **Expérience** la donnée de la famille :

$$(\Omega, \mathcal{F}, X, \mathcal{X}, \mathcal{A}, P_\theta, \theta \in \Theta) \quad \text{où}$$

- $\mathcal{X}, \mathcal{A}$  est l'espace des réalisations de la variable aléatoire  $X$  définie sur  $(\Omega, \mathcal{F})$ .
- $\Theta$  est l'ensemble des **paramètres**.
- $P_\theta$  est une loi de probabilité sur  $(\Omega, \mathcal{F})$ .

Le statisticien fait donc l'hypothèse que son observation  $x$  est la réalisation d'une variable aléatoire (i.e. il existe  $\omega, x = X(\omega)$ ) et qu'il existe  $\theta$  tel que  $\omega$  est tiré selon la loi  $P_\theta$  (la loi de  $X$  est alors  $P_\theta^X$ ).

**Définition 2** Dans une expérience  $\mathcal{E} = (\Omega, \mathcal{F}, X, \mathcal{X}, \mathcal{A}, P_\theta, \theta \in \Theta)$ , on appelle **statistique** toute variable aléatoire de la forme  $T \circ X$  où  $T$  est mesurable de  $(\mathcal{X}, \mathcal{A})$  dans un espace arbitraire muni d'une tribu.

Le statisticien va donc disposer de toutes les "statistiques" comme outil pour "deviner"  $\theta$ .

**Premiers exemples.**

1. Considérons l'exemple du **sondage** : Soit  $N$  et  $n$  fixés. On considère une population de  $N$  éléments qui comprend une proportion  $\theta$  de défectueux. On extrait au hasard  $n$  éléments et on compte le nombre de défectueux parmi cette population extraite. Ce nombre  $x$  est la réalisation d'une variable aléatoire  $X$ . La loi de  $X$  sous  $P_\theta$  est une hypergéométrique que l'on peut écrire sous la forme :

$$P_\theta^X = \sum_{k=0}^n \frac{C_{\theta N}^k C_{(1-\theta)N}^{n-k}}{C_N^n} \delta_k$$

on a biensur  $\Theta = [0, 1]$ ,  $\mathcal{X} = \{0, \dots, n\}$ ,  $\mathcal{A}$  est la tribu des parties de  $\mathcal{X}$ . Par commodité, on pourra prendre  $\Omega = \mathcal{X}$  et  $X$  est alors l'identité.

2. Supposons que l'on observe  $n$  données  $x_1, \dots, x_n$  qui chacune représente une mesure d'une quantité physique  $\mu$ , inconnue que l'on cherche à estimer. Chacune de ces données  $x_i$  est entachée d'une erreur due à la mesure. Faire des statistiques consiste à "modéliser" cette erreur, c'est à dire à considérer par exemple que  $x_i$  peut s'écrire  $\mu + e_i$  où  $e_i$  (l'erreur, qui est tout aussi inconnue que  $\mu$ ) est la **réalisation** d'une variable  $\varepsilon_i$ . De sorte que  $x_i$  est aussi la **réalisation** d'une variable  $X_i = \mu + \varepsilon_i$ .

Il est très important de faire la différence entre les variables que nous considérerons, d'un point de vue théorique pour construire ou valider des procédures et les réalisations de ces variables, qui sont les données numériques que l'on traite par le calcul ou en utilisant des logiciels.

Nous modéliserons ici les erreurs  $\varepsilon_i$  par des variables aléatoires indépendantes, identiquement distribuées de loi  $N(0, \sigma^2)$ , de sorte que  $X_1, \dots, X_n$  sont i.i.d.  $N(\mu, \sigma^2)$ .

On a vu que nous avons en fait souvent pris  $\Omega = \mathcal{X}$ ,  $\mathcal{F} = \mathcal{A}$  en considérant que  $X$  était l'identité. Dans ce cas nous résumerons la donnée du modèle statistique à  $(\mathcal{X}, \mathcal{A}, P_\theta, \theta \in \Theta)$

**Echantillonnage**

**Définition 3** On appelle **modèle d'échantillonnage** associé au modèle  $(\Omega, \mathcal{F}, X, \mathcal{X}, \mathcal{A}, P_\theta, \theta \in \Theta)$  le modèle

$$(\Omega^n, \mathcal{F}_n, X_n, \mathcal{X}^n, \mathcal{A}_n, P_\theta^{\otimes n}, \theta \in \Theta)$$

où  $\mathcal{F}_n$  et  $\mathcal{A}_n$  sont les tribus produit respectivement sur  $\Omega^n$  et  $\mathcal{X}^n$ , et pour tout  $\theta \in \Theta$ ,  $P_\theta^{\otimes n}$  est la probabilité produit de  $n$  copies indépendantes de la loi  $P_\theta$ , notée aussi  $P_\theta^{\otimes n}$ . De plus, si  $\omega^n = (\omega_1, \dots, \omega_n)$  est un élément générique de  $\Omega^n$ ,  $X_n(\omega^n) = (X(\omega_1), \dots, X(\omega_n))$ .

**Exemples d'échantillonnage**

1. En médecine ou en fiabilité on s'intéresse souvent au temps de 'survie' d'un individu ou d'une machine. Prenons le cas des machines : Supposons que nous disposons des temps de panne de  $n$  machines à laver de même marque. On peut faire l'hypothèse que ces machines n'étant pas reliées, leurs pannes sont indépendantes. Il s'agit ensuite de modéliser la loi d'un temps de panne. Plusieurs éventualités sont possibles. Nous allons en envisager deux très différentes.
  - i) Supposons d'abord que l'on fasse l'hypothèse que la machine ne s'use pas : nous avons alors pour tout  $a \leq b$ ,  $t \in \mathbb{R}^+$ ,

$$P(X \in [a+t, b+t] | X \geq t) = P(X \in [a, b] | X \geq 0)$$

On peut montrer que nécessairement, cette loi admet une densité de la forme :

$$f(x) = \lambda \exp -\lambda x.$$

On pourra considérer  $\Phi(x) = P(X > x)$  et montrer que si  $\Phi$  est continue alors le résultat est facile à obtenir.

C'est ce qu'on appelle une loi exponentielle de paramètre  $\lambda > 0$ . Notre modèle est alors un  $n$  échantillon du modèle  $(\mathbb{R}^+, \mathcal{B}(\mathbb{R}^+), X, \mathbb{R}^+, \mathcal{B}(\mathbb{R}^+), P_\lambda, \lambda \in \mathbb{R}_*^+)$ ,  $X$  est l'identité et  $P_\lambda$  la loi exponentielle de paramètre  $\lambda$ .  $(\mathcal{B}(\mathbb{R}^+))$  est la tribu borelienne de  $\mathbb{R}^+$

ii) Supposons maintenant que notre machine ne puisse pas tomber en panne avant un temps connu ( $t = 1$ , par exemple). On prend en compte de cette façon le temps où la machine est sous garantie. On pourra alors considérer un modèle comme ci-dessus mais où  $P_\lambda$  est maintenant la loi de *Pareto* de paramètre  $\lambda > 0$  dont la fonction de répartition est donnée par :

$$G_\lambda(x) = P_\lambda(X \leq x) = 1 - x^{-\lambda}, \text{ si } x \geq 1, 0, \text{ sinon.}$$

*Exercice : Etudier le comportement de cette loi face au vieillissement.*

2. Un autre exemple très classique est le suivant : On observe  $(X_1, \dots, X_n)$   $n$ -variables aléatoires réelles identiquement distribuées de loi  $P$  sur  $\mathbb{R}$ , muni de  $\mathcal{B}(\mathbb{R})$ , sa tribu borelienne et on se propose d'estimer  $P$ , sans autres hypothèses sur  $P$ . Le modèle est alors un modèle d'échantillonnage où l'ensemble des paramètres  $\Theta$  est égal à l'ensemble de toutes les lois de probabilités sur  $\mathbb{R}$ .

**Modèles paramétriques, non-paramétriques** Comme on l'a vu précédemment  $\Theta$  est souvent un sous-ensemble d'un espace  $\mathbb{R}^d$ . Nous dirons quand c'est le cas que le modèle est **paramétrique**.

Le cas ci-dessus où  $\mathcal{X} = \mathbb{R}$ ,  $\mathcal{A}$  est sa tribu borelienne et  $\Theta$  est l'ensemble de toutes les mesures de probabilités sur  $(\mathcal{X}, \mathcal{A})$  est un exemple de modèle **non-paramétrique**.

## 1.2 Modèle Linéaire gaussien

**Définition 4** *Etant donné une matrice  $M$  de dimension  $n \times p$ , On appelle **modèle linéaire gaussien multidimensionnel** associé à la matrice "exogène"  $M$ , une observation  $Y$  dont la loi est  $N_n(M\beta, \sigma^2 I_n)$ .  $\beta$  est un paramètre inconnu de  $\mathbb{R}^p$ .*

*Remarque :* On observe donc, à la fois le vecteur  $Y$  (aléatoire) et la matrice  $M$  supposée déterministe (non aléatoire). On cherche à utiliser cette observation pour tirer des informations sur le paramètre  $\beta$  inconnu. Le modèle précédent peut aussi s'écrire :

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, M = \begin{pmatrix} M_{11} & \dots & M_{1p} \\ & \vdots & \\ M_{n1} & \dots & M_{np} \end{pmatrix}, Y = M\beta + \varepsilon, \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

où les  $\varepsilon_i$  sont i.i.d.  $N(0, \sigma^2)$ .  $\triangle$

Le modèle linéaire gaussien est un modèle statistique, dans lequel, on a

- $\mathcal{X} = \mathbb{R}^n$
- $\mathcal{A}$  est la tribu borelienne de  $\mathbb{R}^n$ .
- $\Theta = \{\theta = (\beta, \sigma^2) \in \mathbb{R}^p \otimes \mathbb{R}_+\}$
- $P_{(\beta, \sigma^2)} = N(M\beta, \sigma^2 I_n)$ .

### 1.2.1 Exemples de modèles linéaires

1. Le modèle précédent de mesure d'une quantités physique est un modèle linéaire.
2. Comparaison de 2 populations de même variance : Supposons que l'on dispose de 2 échantillons  $X_1, \dots, X_n$  i.i.d.  $N(\mu_1, \sigma^2)$  et  $X'_1, \dots, X'_m$  i.i.d.  $N(\mu_2, \sigma^2)$  indépendants. On se demande si ces échantillons sont comparables, autrement dit est-ce que  $\mu_1 = \mu_2$  ?

On concatène les 2 échantillons pour former le vecteur

$$Y = (X_1, \dots, X_n, X'_1, \dots, X'_m)^* = (Y_1, \dots, Y_{m+n})^*$$

Si on considère la matrice  $M$  de taille  $n \times 2$ , telle que

$$M_{11} = \dots = M_{n1} = 1, \quad M_{n+1,1} = \dots = M_{n+m,1} = 0$$

$$M_{12} = \dots = M_{n2} = 0, \quad M_{n+1,2} = \dots = M_{n+m,2} = 1$$

et le vecteur  $\beta = (\mu_1, \mu_2)^*$ , il est facile de mettre notre modèle sous la forme (1.2).

3. Droite de régression. Supposons que l'on sache par des arguments théoriques ( agronomiques, biologiques, économiques, physiques,...) que 2 quantités  $x$  (par exemple le temps) et  $y$  (par exemple la taille d'un animal) sont liées par une équation affine de la forme  $y = ax + b$ , dont on veut identifier les coefficients  $a$  et  $b$ . Une façon de procéder est de mesurer  $y_i$  pour différentes valeurs de  $x_i$  (appelée variable contrôlée ) et de modéliser les erreurs par des  $N(0, \sigma^2)$  indépendantes. On a alors la représentation (1.2), avec

$$M_{11} = x_1, \dots, M_{n1} = x_n,$$

$$M_{12} = 1, \dots, M_{n2} = 1,$$

$$\beta = (a, b)^*$$

Cet exemple peut se généraliser en remplaçant la relation affine par une relation de la forme :

$$y = \sum_{j=0}^p \beta_j f_j(x)$$

Une régression polynomiale s'obtient par exemple en prenant

$$f_0 = 1, \quad f_1(x) = x, \dots, f_p(x) = x^p$$

4. On appelle **Analyse de la variance** (Anova) le cas où la matrice  $M$  est uniquement constituée de 1 et de 0.

Donnons un exemple : Dans des conditions de culture de référence (0), une variété de blé a un rendement moyen de  $\mu$ . On la soumet, dans des parcelles expérimentales à un traitement à 2 facteurs :

1er facteur (par exemple, un engrais) auquel, outre le niveau 0 de référence, on donne 2 niveaux, notés 1 et 2 (par exemple, 2 doses différentes d'engrais).

2eme facteur (par exemple, un niveau d'ensoleillement) auquel on donne soit le niveau de référence 0 soit le niveau 1.

Le modèle de base choisi est le suivant :

$$y = \mu + \alpha_i + \beta_j \tag{1.1}$$

Il est dit additif : Les effets des facteurs s'ajoutent simplement sans interférences.  $\alpha_i$  représente l'effet du 1er facteur au niveau  $i = 0, 1, 2$ ,  $\beta_j$  représente l'effet du 2eme facteur au niveau  $j = 0, 1$ .  $\alpha_0 = \beta_0 = 0$ . Il est clair qu'on aurait pu aussi rajouter "une interaction" de la forme  $\gamma_{ij}$ , mais par souci de simplicité, nous ne l'avons pas fait ici.

Le but est d'obtenir des informations (estimation ou test) sur les  $\alpha_i$  et les  $\beta_j$ . Pour cela, on réalise une expérimentation : On divise un champs en parcelles numérotées (6, dans l'exemple qui suit). Sur chaque parcelle, on applique les facteurs à un niveau prescrit. La description des niveaux affectés aux parcelles s'appelle le plan de l'expérience. Ici, il est donné par le tableau suivant.

Parcelle	1	2	3	4	5	6
Facteur 1	0	1	2	0	1	0
Facteur 2	0	0	0	0	0	1

Si l'on suppose que l'on modélise le rendement sur chaque parcelle par un effet de type (1.1) auquel s'ajoute une erreur  $N(0, \sigma^2)$ , et si l'on suppose les erreurs indépendantes, on obtient une équation du type  $Y = M\beta + \varepsilon$ , où  $Y$  est le vecteur des rendements,  $\varepsilon$  est le vecteur des erreurs,  $\beta = (\mu, \alpha_1, \alpha_2, \beta_1)^*$  et  $M$  est la matrice suivante

$$M = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix}$$

*Exercice* : Sous quelle condition un modèle linéaire est-il un modèle d'échantillonnage?  $\triangle$

### 1.3 Identifiabilité, Domination

**Définition 5** On dit qu'un modèle  $(\mathcal{X}, \mathcal{A}, P_\theta, \theta \in \Theta)$  en position canonique est **identifiable** si l'application de  $\Theta$  dans l'espace des probabilités sur  $(\mathcal{X}, \mathcal{A})$  qui à  $\theta$  associe  $P_\theta$  est injective.

La définition précédente est assez naturelle. On conçoit en effet avoir fort peu de chance d'identifier  $\theta$ , au vu d'une observation  $X$  dont la loi est  $P_\theta^X$  s'il n'y a pas injectivité. Notons qu'un modèle donné n'est pas automatiquement identifiable. Nous en verrons des exemples plus loin, mais il suffit de reprendre l'exemple ci-dessus de l'ANOVA en modèle additif à 2 facteurs. Nous avons fait l'hypothèse  $\alpha_0 = \beta_0 = 0$ . Cette hypothèse avait en fait pour but de rendre le modèle identifiable. Montrer en effet (en exercice) que si les paramètres  $\alpha_0$  et  $\beta_0$  sont inconnus, alors, le modèle précédent n'est pas identifiable.

**Définition 6** On dit qu'un modèle en position canonique  $(\mathcal{X}, \mathcal{A}, P_\theta, \theta \in \Theta)$  est **dominé** s'il existe une mesure  $\mu$  sur  $(\mathcal{X}, \mathcal{A})$ , positive,  $\sigma$ -finie telle que pour tout  $\theta \in \Theta$ ,  $P_\theta$  est dominée par  $\mu$ .  $\mu$  est appelée *mesure dominante* du modèle.

Si la définition précédente était assez naturelle, celle-ci l'est un peu moins. Disons seulement ici que sans cette propriété, beaucoup de résultats seront faux. Nous verrons ( et heureusement ) que beaucoup de modèles 'standards' sont dominés. Ce n'est toutefois pas toujours le cas. Le premier exemple ci-dessous montre que l'ensemble de toutes les lois sur  $\mathbb{R}, \mathcal{B}(\mathbb{R})$  n'est pas dominé.

*Remarques :*

1. Rappelons les notions utilisées dans la définition précédente : On dit qu'une mesure positive  $\mu$  sur  $(\mathcal{X}, \mathcal{A})$  est  $\sigma$ -finie si il existe une suite  $A_n$  d'ensembles de  $\mathcal{A}$  telle que  $\mathcal{X} = \cup_n A_n$  et  $\mu(A_n) < \infty, \forall n$ .
2. On dit que la mesure  $\mu$  domine  $P_\theta$  (noté  $P_\theta \ll \mu$ ), si pour tout  $A$  de  $\mathcal{A}$ ,  $\mu(A) = 0$  entraîne  $P_\theta(A) = 0$ .
3. La domination est une relation transitive. On en déduit donc que si un modèle est dominé, il l'est par une infinité de mesures positives. (En particulier toutes les mesures qui dominent  $\mu$ .)
4.  $P_\theta \ll \mu$  équivaut, d'après le théorème de Radon-Nykodym à l'existence d'une densité de  $P_\theta$  par rapport à  $\mu$ ,  $\frac{dP_\theta}{d\mu}$ , définie  $\mu$  presque sûrement, telle que pour toute fonction  $F$ ,  $P_\theta$ -intégrable on a :

$$\int_{\mathcal{X}} F(x) dP_\theta = \int_{\mathcal{X}} F(x) \frac{dP_\theta}{d\mu} d\mu$$

5. Dans un modèle dominé, on peut définir des fonctions, appelées fonctions de vraisemblance, de  $\mathcal{X} \otimes \Theta$  dans  $\mathbb{R}_+$  qui à  $(x, \theta)$  associent

$$p(x, \theta) = \frac{dP_\theta}{d\mu}(x)$$

$\triangle$

### Identifiabilité, domination : Exemples

1. Si  $\delta_x$  désigne la masse de Dirac au point  $x$  et  $\mathcal{B}(E)$  la tribu borelienne de l'espace topologique  $E$ , on peut remarquer que

$$(\mathbb{R}, \mathcal{B}(\mathbb{R}), \delta_\theta, \Theta)$$

est dominé si et seulement si  $\Theta$  est un ensemble dénombrable de  $\mathbb{R}$ .

En effet si  $\Theta = \{\theta_i, i \in \mathbb{N}^*\}$  est dénombrable, alors  $\mu = \sum_{\Theta} \delta_\theta$  domine le modèle. (On montre que cette mesure est  $\sigma$ -finie en décomposant  $\mathbb{R}$  en l'union des  $A_i = \{\theta_i\}, i \in \mathbb{N}^*$ , de  $\mu$ - mesure positive, et  $A_0 = \mathbb{R} \setminus (\cup_{i \in \mathbb{N}^*} \{\theta_i\})$  de  $\mu$ - mesure nulle.)

Si  $\Theta$  n'est pas dénombrable, et si  $\mu$  est une mesure dominante du modèle, alors si on considère  $\frac{d\delta_\theta}{d\mu}$ , la densité de  $\delta_\theta$  par rapport à  $\mu$ , on a

$$1 = \delta_\theta(\theta) = \frac{d\delta_\theta}{d\mu}(\theta) \mu(\{\theta\})$$

Ceci entraîne que  $\forall \theta \in \Theta, \mu(\{\theta\}) \neq 0$ .

Nous avons le lemme suivant :

**Lemme 1** *Si  $B$  est tel que  $\mu(B) < +\infty$  et  $\mu\{x\} > 0, \forall x \in B$ ,  $B$  est nécessairement un ensemble dénombrable.*

*En effet, soit  $A = \{x, \mu\{x\} > 0\}$ ,  $A = \cup_{k>0} E_k$ ,  $E_k = \{x, \mu\{x\} \geq \frac{1}{k}\}$ . On a  $\mu(E_k) \geq \frac{1}{k} \text{card}\{E_k\}$ . Et donc,  $\mu(E_k) < \infty$  implique que  $E_k$  est un ensemble fini. Donc si  $B \subset A$  est tel que  $\mu(B) < \infty$ , nécessairement  $B$  est union dénombrable d'ensembles finis donc dénombrable.*

On en déduit que  $\mu$  n'est pas donc pas  $\sigma$ -finie. En effet, si on considère un recouvrement de  $\mathbb{R}$  par des ensembles mesurables  $\mathcal{A}_n$  tels que  $\mu(\mathcal{A}_n) < \infty$ , on en déduit un recouvrement de  $\Theta$  par une suite d'ensembles de mesure finie et chargeant chaque points, donc dénombrables en utilisant le lemme précédent. Ceci conduit au fait que  $\Theta \subset \mathbb{R}$  est dénombrable. Une conséquence, plus intéressante en pratique est que le modèle non paramétrique cité plus haut de toutes les probabilités sur  $\mathbb{R}$  n'est pas dominé.

2. Le modèle linéaire gaussien  $Y = M\beta + \varepsilon$  est un modèle dominé (on peut prendre  $\mu$  la mesure de Lebesgue sur  $\mathbb{R}^n$ ). Il est identifiable si et seulement si  $M$  est injective (exercice). Une vraisemblance est :

$$\frac{dP_\theta}{d\mu}(y) = \frac{\exp\left\{-\frac{\|y - M\beta\|^2}{2\sigma^2}\right\}}{(\sqrt{2\pi}\sigma)^n}$$

3. Considérons l'exemple du sondage : Soit  $N$  et  $n$  fixés. On considère une population de  $N$  éléments qui comprend une proportion  $\theta$  de defectueux. On extrait au hasard  $n$  éléments et on compte le nombre de defectueux parmi cette population extraite. La loi  $P_\theta$  de ce nombre de defectueux est une hypergéométrique que l'on peut écrire sous la forme :

$$P_\theta = \sum_{k=0}^n \frac{C_{\theta N}^k C_{(1-\theta)N}^{n-k}}{C_N^n} \delta_k$$

Ce modèle est dominé par la mesure  $\mu = \sum_{k=0}^n \delta_k$  et une vraisemblance est

$$\frac{dP_\theta}{d\mu}(k) = \frac{C_{\theta N}^k C_{(1-\theta)N}^{n-k}}{C_N^n}$$

Montrer en exercice que ce modèle est identifiable et que ça n'est pas un modèle d'échantillonnage.

4. Montrer en exercice qu'un modèle est identifiable (resp. dominé) si et seulement si le modèle échantillonné l'est.
5. On observe  $n$  sites de pontes. Pour chaque site, le nombre d'oeufs suit une loi de Poisson de paramètre  $\lambda$ . Chaque oeuf a une probabilité  $p$  de donner un insecte. On considère que les oeufs sont indépendants entre eux, indépendants du nombre  $X$  et que les sites de pontes sont aussi indépendants entre eux. On peut faire 2 types d'observations
- Soit on observe  $X$  le nombre d'oeufs et  $Y$  le nombre d'insectes (observation complète), sur chaque site de ponte.
  - Soit on observe seulement  $Y$  le nombre d'insectes (observation partielle), également sur chaque site de ponte.

On a  $\Theta = \{(\lambda, p) \in \mathbb{R}_+^* \otimes [0, 1]\}$

- Dans le premier cas, l'espace  $\mathcal{X} = (\Lambda_1, \dots, \Lambda_n)$  avec  $\Lambda_i = \mathbb{N} \otimes \mathbb{N}$ , muni de la tribu des parties.  $P_\theta^{\otimes n}$  est la loi de  $n$  copies indépendantes de la loi  $P_\theta$  du couple  $(X, Y)$ .

Calculons cette loi : On a

$$\begin{aligned} P_\theta(Y = y | X = x) &= C_x^y p^y (1-p)^{x-y} \mathbf{1}_{0 \leq y \leq x}, & P_\theta(X = x) &= e^{-\lambda} \frac{\lambda^x}{x!} \\ P_\theta(Y = y, X = x) &= C_x^y p^y (1-p)^{x-y} \mathbf{1}_{0 \leq y \leq x} e^{-\lambda} \frac{\lambda^x}{x!} \end{aligned}$$

On obtient donc,

$$P_\theta^{\otimes n}((x_1, y_1), \dots, (x_n, y_n)) = \prod_{i=1}^n C_{x_i}^{y_i} p^{y_i} (1-p)^{x_i - y_i} \mathbf{1}_{0 \leq y_i \leq x_i} e^{-\lambda} \frac{\lambda^{x_i}}{x_i!}$$

L'exercice ci-dessus montre qu'il nous suffit d'étudier le cas  $n = 1$ . Le modèle est dominé, nous laissons la démonstration au lecteur. Pour vérifier que le modèle est identifiable, prenons  $\lambda, p, \lambda', p'$  et supposons que

$$C_x^y p^y (1-p)^{x-y} e^{-\lambda} \frac{\lambda^x}{x!} = C_x^y p'^y (1-p')^{x-y} e^{-\lambda'} \frac{\lambda'^x}{x!} \quad \forall x \geq 0, 0 \leq y \leq x \quad (1.2)$$

Supposons d'abord que  $p$  et  $p'$  sont dans  $]0, 1[$ . Alors l'équation (1.2) devient :

$$\left(\frac{p(1-p')}{p'(1-p)}\right)^y \left(\frac{\lambda(1-p)}{\lambda'(1-p')}\right)^x = \exp\{(\lambda - \lambda')\}$$

que l'on peut encore écrire  $A^y B^x = \exp\{(\lambda - \lambda')\} \quad \forall x \geq 0, 0 \leq y \leq x$ . En passant au logarithme, on vérifie aisément que ceci n'est possible que si  $A = B = 1, \lambda' = \lambda$ .

Remarquons maintenant que si  $p = 0$ , si on se place en  $y = 0$ , (1.2) s'écrit :

$$e^{\lambda - \lambda'} \left[\frac{(1-p')\lambda'}{\lambda}\right]^x = 1, \quad \forall x \geq 0$$

En passant au logarithme, on vérifie aisément que ceci implique  $p' = 0, \lambda = \lambda'$ .

On fait un raisonnement analogue pour le cas  $p = 1$ . On en conclut que le modèle est aussi identifiable pour l'ensemble des paramètres  $\Theta$ .

6. Si nous considérons maintenant le cas où on observe seulement  $Y$  l'espace est maintenant  $\mathcal{X} = (\Lambda_1, \dots, \Lambda_n)$  avec  $\Lambda_i = \mathbb{N}$ , muni de la tribu des parties.  $Q_\theta^{\otimes n}$  est la loi de  $n$  copies indépendantes de la loi  $Q_\theta$  de la variable  $Y$ . On a

$$\begin{aligned} Q_\theta(Y = y) &= \sum_{y \leq x} e^{-\lambda} \frac{\lambda^x}{x!} C_x^y p^y (1-p)^{x-y} 1_{0 \leq y} \\ &= \frac{e^{-\lambda} \lambda^y p^y}{y!} \sum_{y \leq x} \frac{((1-p)\lambda)^{(x-y)}}{(x-y)!} \\ &= \frac{e^{-\lambda p} (p\lambda)^y}{y!} \end{aligned}$$

On en déduit que  $Q_\theta$  est une loi de Poisson de paramètre  $\lambda p$ . Le modèle est clairement non identifiable.

*Remarque :* On peut retrouver la loi de  $Y$  en remarquant que

$$Y = \sum_{i=1}^X Z_i$$

où les  $Z_i$  sont i.i.d. de loi de Bernoulli de paramètre  $p$  (i.e.  $P(Z_i = 1) = p = 1 - P(Z_i = 0)$ ) et indépendantes de  $X$ . Calculer la fonction génératrice  $\mathbb{E}s^Y$  pour  $s$  réel de module strictement inférieur à 1.

△

## 1.4 Structure des modèles dominés.

Le but de cette section est d'abord de montrer que parmi les mesures dominantes, certaines peuvent avoir des propriétés particulières. Nous allons ensuite nous attacher à donner une formulation mathématique à l'observation que l'on a pu faire au travers des exemples précédents : une famille dominée est une famille 'pas trop grande' en un sens à déterminer.

### 1.4.1 Dominante Privilégiée

Nous avons vu que les mesures dominantes ne sont pas uniques. Toutefois certaines sont plus intéressantes que d'autres.

**Proposition 1** *Si un modèle  $(\mathcal{X}, \mathcal{A}, P_\theta, \theta \in \Theta)$  est dominé par une mesure  $\mu$ , il est dominé par une probabilité  $Q$ , vérifiant de plus  $Q(A) = 0 \iff P_\theta(A) = 0, \forall \theta \in \Theta$ . Une telle probabilité est appelée **dominante privilégiée** du modèle.*

*Remarque :* Pas plus qu'une mesure dominante, une probabilité dominante privilégiée n'est unique. N'importe qu'elle probabilité équivalente à la probabilité de départ convient.  $\triangle$

*Preuve de la Proposition :*

La démonstration de la proposition repose sur le lemme suivant :

**Lemme 1** *Toute mesure  $\sigma$ -finie est dominée par une probabilité.*

*Démonstration du lemme :*

Nous savons qu'il existe des ensembles mesurables,  $A_n, n \in \mathbb{N}$  tels que  $\mathcal{X} = \cup_n A_n$  et  $\mu(A_n) < \infty, \forall n \in \mathbb{N}$ . Soit  $\lambda_n$  une suite de réels tels que  $\lambda_n \in [0, 1], \sum \lambda_n = 1$  et  $\lambda_n > 0 \iff \mu(A_n) > 0$ . Soit  $P$  la mesure de densité par rapport à  $\mu$  :

$$\frac{dP}{d\mu}(x) = \sum_{n \in \mathbb{N}} \frac{\lambda_n}{\mu(A_n)} 1_{A_n}(x)$$

Il est clair que  $P(A) = \sum_{n \in \mathbb{N}} \frac{\lambda_n \mu(A \cap A_n)}{\mu(A_n)}$  donc  $P$  est une probabilité. Par ailleurs  $P$  est équivalente à  $\mu$  : En effet

$$P(A) = 0 \iff \mu(A \cap A_n) = 0, \forall n \iff \mu(A) = 0.$$

■

Passons maintenant à la démonstration de la proposition : En utilisant le lemme, considérons  $P$  probabilité, équivalente à  $\mu$ .  $P$  domine toutes les  $P_\theta$  par transitivité, mais il va nous falloir travailler plus pour fabriquer une dominante privilégiée  $Q$ . Posons  $F_\theta = \frac{dP_\theta}{dP}$  et  $A_\theta = \{F_\theta > 0\} \in \mathcal{A}$ . Considérons  $\mathcal{C} \subset \mathcal{A}$ , l'ensemble des réunions dénombrables d'éléments  $A_\theta$  et

$$M = \sup_{C \in \mathcal{C}} P(C).$$

On vérifie que  $M \leq 1$ , et que  $M$  est atteinte : Il existe une suite  $C_n$ , telle que  $P(C_n) \geq M - n^{-1}$ . Posons  $C^* = \cup_n C_n$ ,  $C^*$  appartient à  $\mathcal{C}$  puisque cet ensemble est stable par réunion dénombrable. On a  $M \leq P(C^*) \leq M$ . Par ailleurs, toujours par définition de  $\mathcal{C}$ , il existe une suite  $\{\theta_j, j \in \mathbb{N}\}$  telle que  $C^* = \cup_j A_{\theta_j}$ . Soit  $\lambda_n$  une suite de réels,  $\lambda_n \in ]0, 1], \sum \lambda_n = 1$ , Définissons la mesure  $Q$  dont la densité par rapport à  $P$  est

$$\frac{dQ}{dP}(x) = \sum_{j \in \mathbb{N}} \lambda_j F_{\theta_j}(x).$$

Il est évident  $Q(A) = \sum_{j \in \mathbb{N}} \lambda_j P_{\theta_j}(A)$  donc  $Q$  est une probabilité et si  $A \in \mathcal{A}$  est tel que  $P_\theta(A) = 0, \forall \theta \in \Theta$  alors  $Q(A) = 0$ .

Le point difficile est de démontrer que  $Q$  est dominante. Remarquons d'abord que  $C^*$  a la propriété fondamentale suivante (support maximal) :

$$\forall \theta \in \Theta, P(A_\theta) = P(A_\theta \cap C^*) \quad (1.3)$$

En effet, sinon,  $P(A_\theta \cap (C^*)^c)$  serait strictement positif et  $P(A_\theta \cup C^*) = P(C^*) + P(A_\theta \cap (C^*)^c)$  serait strictement plus grand que  $M$ .

Considérons  $A$  tel que  $Q(A) = 0$ , on a, bien entendu  $P_{\theta_j}(A) = 0 \forall j \in \mathbb{N}^*$ , mais montrons qu'on a bien  $P_\theta(A) = 0 \forall \theta \in \Theta$  :

$$P_\theta(A) = \int_A F_\theta dP = \int_{A \cap A_\theta} F_\theta dP = \int_{A \cap A_\theta \cap C^*} F_\theta dP$$

En effet, à cause de (1.3),  $P(A_\theta \cap (C^*)^c) = 0$  et donc  $P_\theta(A_\theta \cap (C^*)^c) = 0$  par domination. Par conséquent,  $P_\theta(A \cap A_\theta \cap (C^*)^c) = 0$ . On en déduit,

$$\begin{aligned} P_\theta(A) &= \int_{A \cap A_\theta \cap (\cup_j A_{\theta_j})} F_\theta dP \\ &\leq \sum_j \int_{A \cap A_\theta \cap A_{\theta_j}} F_\theta dP \\ &= \sum_j \int_{A \cap A_\theta \cap A_{\theta_j}} \frac{F_\theta}{F_{\theta_j}} F_{\theta_j} dP \\ &= \sum_j \int_{A \cap A_\theta \cap A_{\theta_j}} \frac{F_\theta}{F_{\theta_j}} dP_{\theta_j} \end{aligned}$$

Or, par définition  $Q(A) = 0 \iff P_{\theta_j}(A) = 0 \forall j \in \mathbb{N}$ . En conséquence la somme précédente est nulle. ■

#### 1.4.2 Distance en variation et Modèles dominés

**Définition 7** Soit  $(\mathcal{X}, \mathcal{A})$  un ensemble et sa tribu et désignons par  $\mathcal{P}$  l'ensemble de toutes les probabilités sur  $(\mathcal{X}, \mathcal{A})$ . On peut munir  $\mathcal{P}$  de la **distance en variation** définie comme suit :

$$d(P, Q) = \sup_{A \in \mathcal{A}} |P(A) - Q(A)|$$

*Remarque* : Il est clair que  $d$  est une distance :

- $d(P, Q) = d(Q, P)$ .
- $d(P, Q) = 0 \iff P = Q$ .
- $d(P, Q) \leq d(P, R) + d(R, Q)$ .

△

**Lemme 2** Si  $S$  domine  $P$  et  $Q$ , soit  $p = \frac{dP}{dS}$ ,  $q = \frac{dQ}{dS}$ , alors

$$d(P, Q) = \frac{1}{2} \int |p - q| dS = \int_{p > q} (p - q) dS = \int_{q > p} (q - p) dS$$

*Démonstration du lemme*

- Remarquons tout d'abord qu'une telle mesure existe toujours : Il suffit de prendre  $S = P + Q$ .

- Nous avons  $0 = \int (p - q)dS = \int_{p>q}(p - q)dS - \int_{q>p}(q - p)dS$ .  
De plus,  $\int |p - q|dS = \int_{p>q}(p - q)dS + \int_{q>p}(q - p)dS$ .
- Maintenant,  $P(A) - Q(A) = Q(A^C) - P(A^C)$ , ce qui implique :  
 $|P(A) - Q(A)| = \frac{1}{2}\{|P(A) - Q(A)| + |Q(A^C) - P(A^C)|\}$   
 $= \frac{1}{2}\{|\int_A dP - dQ| + |\int_{A^C} dP - dQ|\}$ .  
On en déduit :  $|P(A) - Q(A)| \leq \frac{1}{2}\{\int_A |p - q|dS + \int_{A^C} |p - q|dS\}$ , soit,  
 $d(P, Q) \leq \frac{1}{2} \int |p - q|dS$ .
- Soit maintenant  $A = \{p > q\}$ , on a  $d(P, Q) \geq |P(A) - Q(A)| = \int_{p>q}(p - q)dS$ .

■

**Proposition 2** *On a :*

- $0 \leq d(P, Q) \leq 1$ .
- $d(P, Q) = 1 \iff P$  et  $Q$  sont étrangères. (i.e. il existe  $A \in \mathcal{A}$ , tel que  $P(A) = 1, Q(A) = 0$ .)

*Démonstration de la Proposition*

- D'abord on a  $0 \leq d(P, Q) = \frac{1}{2} \int |p - q|dS \leq \frac{1}{2} \int (p + q)dS$ .
- Supposons maintenant :  
 $1 = d(P, Q) = \int_{p \geq q}(p - q)dS = \int_{p \geq q} pdS - \int_{p \geq q} qdS$ .  
Les 2 quantités de cette différence étant comprises entre 0 et 1, ceci n'est possible que si  $\int_{p \geq q} pdS = 1$  et  $\int_{p \geq q} qdS = 0$ . Il suffit alors de prendre  $A = \{p \geq q\}$ , et de vérifier qu'on a bien  $P(A) = 1, Q(A) = 0$ .

**Théorème 1** *Soit  $(\mathcal{X}, \mathcal{A}, P_\theta, \theta \in \Theta)$  un modèle statistique.*

- *Si la famille  $\{P_\theta, \theta \in \Theta\}$  est séparable pour la distance en variation, alors le modèle est dominé.*
- *Si la tribu  $\mathcal{A}$  est dénombrablement engendrée, la réciproque est vraie.*

*Remarques :*

1. Un ensemble est dit séparable pour une métrique, s'il contient un ensemble dénombrable dense pour cette métrique.
2. Une tribu est dite dénombrablement engendrée si elle s'écrit  $\sigma(A_n, n \in \mathbb{N})$ .  $\triangle$

**Corollaire 1** *Si  $\Theta$  peut être muni d'une métrique pour laquelle il est séparable, et si l'application  $\phi$  de  $\Theta$  dans  $\mathcal{P}$ , muni de la distance en variation*

$$\theta \in \Theta \xrightarrow{\phi} P_\theta \in \mathcal{P}$$

*est continue, alors le modèle est dominé.*

La démonstration du corollaire est triviale si on remarque que l'image d'un espace séparable par une application continue est séparable.

**Démonstration du Théorème**

- Supposons donc  $(P_\theta, \theta \in \Theta, d)$  séparable. Cela signifie qu'il existe une famille  $(\theta_n)_{n \in \mathbb{N}^*}$  dense pour la métrique  $d$ . Considérons la mesure de probabilité

$$S = \sum_{n \in \mathbb{N}^*} \frac{1}{2^n} P_{\theta_n}$$

et vérifions que  $S$  est une dominante privilégiée.

En effet, soit  $A \in \mathcal{A} / S(A) = 0$ , on a  $P_{\theta_n}(A) = 0 \forall n \in \mathbb{N}^*$  par construction de  $S$ .

Montrons que ceci implique que  $P_\theta(A) = 0 \forall \theta \in \Theta$  :

En effet, sinon soit  $\theta$  tel que  $P_\theta(A) > a > 0$ . Par densité, il existe,  $\theta_n$  tel que

$d(P_\theta, P_{\theta_n}) \leq a/2$ , mais ceci implique que  $|P_\theta(A) - P_{\theta_n}(A)| \leq a/2$ . Nous aboutissons à une contradiction.

- Pour démontrer la réciproque, nous laissons en exercice, le résultat suivant (la construction est usuelle et repose sur la considération des fonctions  $\{\sum_{i=1}^N \alpha_i 1_{A_i}, \alpha_i \in \mathbb{Q}, N \in \mathbb{N}\}$  :

*Si  $\mathcal{A}$  est dénombrablement engendrée et  $\mu$   $\sigma$ -finie alors  $\mathbb{L}_1(\mu) = \{f / \int |f| d\mu < \infty\}$ , est un ensemble séparable.*

Supposons que  $S$  domine la famille  $\{P_\theta, \theta \in \Theta\}$ , alors  $\{P_\theta, \theta \in \Theta\}$  s'injecte naturellement dans  $\mathbb{L}_1(S)$ . En effet si on note  $p_\theta = \frac{dP_\theta}{dS}$ , on a  $2d(P_\theta, P_{\theta'}) = \int |p_\theta - p_{\theta'}| dS$ . Or cette dernière quantité est exactement la norme dans  $\mathbb{L}_1(S)$  de  $p_\theta - p_{\theta'}$ . La séparabilité de  $\{P_\theta, \theta \in \Theta\}$  se déduit alors de celle de  $\mathbb{L}_1(S)$ .

## 1.5 Introduction à l'exhaustivité.

Ce problème concerne les expériences où l'on dispose de beaucoup de données, et que (pour des problèmes de stockage par exemple), on cherche à réduire la taille de ces données. C'est souvent le cas par exemple en traitement de l'image : Une image peut être la donnée de  $256 \times 256$  niveaux de gris par "pixels" (i.e. picture elements : éléments d'image). Cette donnée définit une qualité d'image. Toutefois, si l'on désire transmettre cette image rapidement, il peut être utile de résumer cette donnée. Bien entendu, on souhaite généralement que cette réduction se fasse avec le moins de perte possible au niveau de la qualité de l'image. La notion de perte de qualité est relativement difficile à définir de façon quantitative. Nous allons étudier maintenant la notion d'exhaustivité, qui correspond à une réduction de la taille des données, sans perte d'information statistique.

### 1.5.1 Sous-Expérience, perte d'information.

Sans aller plus loin dans la formalisation à ce niveau, disons que nous allons considérer que 'l'information' que nous amène une expérience est l'ensemble des 'statistiques' (cf définition) qu'elle permet de construire.

**Définition 8**  $\mathcal{E} = (\Omega, \mathcal{F}, X, \mathcal{X}, \mathcal{A}, P_\theta, \theta \in \Theta)$  étant une expérience, et  $T$  une application mesurable de  $(\mathcal{X}, \mathcal{A})$  dans  $(\Pi, \mathcal{T})$ , on appelle **sous expérience** associée à  $T$ , l'expérience :

$$\mathcal{E}^T = (\Omega, \mathcal{F}, T \circ X, \Pi, \mathcal{T}, P_\theta, \theta \in \Theta)$$

*Remarques :*

1. Un exemple consiste dans le cas du modèle d'échantillonnage où l'on observe  $(X_1, \dots, X_n)$  à prendre  $T$  telle que  $T(x_1, \dots, x_n) = x_1$ . Les statistiques de l'expérience  $\mathcal{E}$  sont calculées à partir de  $(X_1, \dots, X_n)$ . En revanche, celles de la sous expérience  $\mathcal{E}^T$  ne sont calculées qu'à partir de  $X_1$ , de sorte que la moyenne  $\bar{X}$  par exemple n'est plus accessible à partir de  $\mathcal{E}^T$ . On voit qu'il peut en résulter une perte d'information et des pertes de précisions statistiques importantes.
2. Dans le paragraphe précédent nous avons rencontré un problème de sous expérimentation qui se pose naturellement dans l'exemple des sites de ponte.

3. L'exemple suivant est aussi très important en statistique. Il s'agit de la **réduction de données en classes**. L'expérience originale est un échantillonnage de taille  $n$ , chaque variable  $X_i$  étant à valeurs dans un ensemble  $X, \mathcal{F}$ . Soit  $A_1, \dots, A_k$  une partition mesurable de  $X, \mathcal{F}$ . Supposons maintenant qu'au lieu d'observer chaque  $X_i$  on sous expérimente en n'observant que la classe  $A_l$  dans laquelle il est tombé. i.e. on observe  $T(X_i) = \sum_{l=1}^k I\{X_i \in A_l\}$ . On peut encore réduire l'information en n'observant que  $(N_1, \dots, N_k)$ , où  $N_l = \sum_{i=1}^n I\{X_i \in A_l\}$  compte le nombre de  $X_i$  qui sont tombés dans chaque classe.
4. La question fondamentale de l'exhaustivité est la suivante : on dira que  $T$  est exhaustive si l'expérience  $\mathcal{E}^T$  est aussi informative que  $\mathcal{E}$ . On peut alors naturellement se demander à quelles conditions sur  $T$  une telle propriété est réalisée.

Il y a un cas où le problème est clair : S'il existe  $U$ , mesurable de  $(\Pi, \mathcal{T})$  dans  $(\mathcal{X}, \mathcal{A})$  tel que  $U(T(x)) = x, \forall x \in \mathcal{X}$ . Dans ce cas,  $\mathcal{E}^T$  est une sous expérience de  $\mathcal{E}$ , mais  $\mathcal{E} = (\mathcal{E}^T)^U$  est aussi une sous expérience de  $\mathcal{E}^T$ . Les 2 expériences sont donc clairement équivalentes au sens où elles permettent de calculer les mêmes statistiques.  $\Delta$

Néanmoins, il existe des cas intéressants où on n'a pas l'existence de  $U$ , mesurable de  $(\Pi, \mathcal{T})$  dans  $(\mathcal{X}, \mathcal{A})$  tel que  $U(T(x)) = x, \forall x \in \mathcal{X}$ , et où on n'a pas perte d'information statistique.

Étudions l'exemple important suivant :

Rappelons que l'on appelle **modèle multinomial** le modèle où

- l'observation  $X$  est à valeurs dans  $\{(n_1, \dots, n_k), n_l \in \mathbb{N}, \sum_{l=1}^k n_l = n\}$  muni de la tribu de ses parties.
- $\theta = (\theta_1, \dots, \theta_k) \in \Theta = \{(\theta_1, \dots, \theta_k), \theta_l \in [0, 1], \sum_{l=1}^k \theta_l = 1\}$
- $P_\theta\{(n_1, \dots, n_k)\} = \frac{n!}{n_1! \dots n_k!} \theta_1^{n_1} \dots \theta_k^{n_k}$ .

Ce modèle provient généralement comme on l'a indiqué dans l'exemple plus haut d'une sous expérimentation qui n'observe pas quels individus sont tombés dans telle classe, mais ne fait que les compter. Nous allons montrer qu'en fait ce genre de sous expérimentation ne réduit pas l'information statistique. Plus précisément, prenons pour simplifier le cas où le nombre de classes  $k = 2$ .

Soit  $(X_1, \dots, X_n)$  un n-échantillon de la loi de Bernoulli de paramètre  $\theta \in [0, 1]$ . Nous allons considérer que 2 types d'observation sont possibles :

1. Soit on observe effectivement  $(X_1, \dots, X_n)$ , et on a

$$\mathcal{E} = (\{0, 1\}^n, \mathcal{A}_n, B(\theta)^{\otimes n}, \theta \in [0, 1]).$$

2. Soit on observe seulement un compteur  $T = \sum_{i=1}^n X_i$  et on a la sous-expérience

$$\mathcal{E}^T = (\{0, \dots, n\}, \mathcal{B}_n, \text{Bin}(n, \theta), \theta \in [0, 1]),$$

où  $\text{Bin}(n, \theta)$  est la loi binomiale (i.e.  $P_\theta(k) = C_n^k \theta^k (1 - \theta)^{n-k} \mathbf{1}_{0 \leq k \leq n}$ ).

$\mathcal{E}^T$  est clairement une sous expérience de  $\mathcal{E}$  mais l'application  $T$  n'a clairement pas d'inverse. Nous allons maintenant montrer que s'il ne nous est pas possible de reconstruire  $(X_1, \dots, X_n)$  à partir de  $T$ , en revanche, il est possible de reconstruire à partir de  $T$ .  $(Y_1, \dots, Y_n)$  de même **loi que**  $(X_1, \dots, X_n)$ . C'est à dire qu'il nous est possible de reconstruire l'expérience  $\mathcal{E}$ .

Considérons la procédure suivante :

- A partir  $T$ , nous construisons  $(Y_1, \dots, Y_n) \in \{0, 1\}^n$  tel que :
- Conditionnellement à  $T = k$ , on tire au hasard (uniformément) un ensemble de  $k$  éléments parmi  $n$ . On obtient ainsi l'ensemble aléatoire  $A_k$ .
- On définit  $Y_i = 1$  si  $i \in A_k, Y_i = 0$  sinon.

Quelle est la loi de  $(Y_1, \dots, Y_n)$ ? Soit  $(\varepsilon_1, \dots, \varepsilon_n) \in \{0, 1\}^n$ , on a

$$\begin{aligned} & P_\theta(Y_1 = \varepsilon_1 \cap \dots \cap Y_n = \varepsilon_n) \\ &= \sum_{k=0}^n \mathbb{1}_{\{\sum_i \varepsilon_i = k\}} P_\theta(Y_1 = \varepsilon_1 \cap \dots \cap Y_n = \varepsilon_n | T = k) P_\theta(T = k) \\ &= \sum_{k=0}^n \mathbb{1}_{\{\sum_i \varepsilon_i = k\}} \frac{1}{C_n^k} C_n^k \theta^k (1 - \theta)^{n-k} \end{aligned}$$

Il est facile de vérifier que

$$P_\theta(Y_1 = \varepsilon_1 \cap \dots \cap Y_n = \varepsilon_n) = P_\theta(X_1 = \varepsilon_1 \cap \dots \cap X_n = \varepsilon_n)$$

Par cette procédure nous avons reconstruit non pas  $(X_1, \dots, X_n)$  à partir de  $T$ , mais  $(Y_1, \dots, Y_n)$ , qui a la même loi. En d'autres termes, à partir de l'expérience  $\mathcal{E}^T$ , on a reconstruit l'expérience  $\mathcal{E}$ .

*Remarque :* Il est fondamental de remarquer ici que cette reconstruction n'a été possible en connaissant  $T$ , mais **sans connaître**  $\theta$  parce que

$$P_\theta(Y_1 = \varepsilon_1 \cap \dots \cap Y_n = \varepsilon_n | T = k) = \frac{1}{C_n^k} \quad (1.4)$$

ne dépend pas de  $\theta$ .  $\triangle$

La définition classique de l'exhaustivité dans un modèle dominé est la suivante :

**Définition 9** *Supposons que le modèle  $\mathcal{E} = (\Omega, \mathcal{F}, X, \mathcal{X}, \mathcal{A}, P_\theta, \theta \in \Theta)$  est dominé.  $T$  étant une application mesurable de  $(\mathcal{X}, \mathcal{A})$  dans  $(\Pi, \mathcal{T})$ , on dira que  $T$  est exhaustive si et seulement si pour tout  $A$  dans  $\mathcal{A}$ , pour tout  $B$  dans  $\mathcal{T}$ , pour tous  $\theta$  et  $\theta'$  dans  $\Theta$ ,*

$$P_\theta(X \in A | T \in B) = P_{\theta'}(X \in A | T \in B). \quad (1.5)$$

Il est clair que vérifier (1.5) dans le modèle précédent équivaut à vérifier (1.4), qui s'est avérée fondamentale dans notre construction. Le lien dans un cadre général entre cette condition et la reconstruction de l'expérience originelle à partir de la sous-expérience est du à LeCam <sup>1, 2, 3</sup>. C'est un travail assez complexe que nous n'aborderons pas ici.

<sup>1</sup>Le Cam, Lucien and Yang, Grace Lo, "Asymptotics in statistics. Some basic concepts", Springer-Verlag, New York, 1990

<sup>2</sup>Strasser, Helmut, Mathematical theory of statistics, Walter de Gruyter & Co., Berlin, 1985

<sup>3</sup>Genon-Catalot, Valentine and Picard, Dominique, Éléments de statistique asymptotique, Springer-Verlag, Paris, 1993,

## Chapitre 2

# VARIABLES ALEATOIRES GAUSSIENNES.

La statistique des variables gaussiennes est une partie fondatrice et très illustrante de la théorie statistique. Nous y consacrerons une partie importante de ce cours. Dans ce but, nous allons, dans ce chapitre rappeler les propriétés des variables et des vecteurs gaussiens dont nous aurons besoin ultérieurement.

### 2.1 Rappel sur les Gaussiennes réelles

**Définition 10**  $X$  est une variable gaussienne standard (centrée, réduite) si et seulement si sa loi admet la densité

$$\varphi(x) = \frac{1}{(2\pi)^{1/2}} \exp \frac{-x^2}{2}$$

par rapport à la mesure de Lebesgue sur  $\mathbb{R}$ .

Les propriétés suivantes sont laissées en exercice :

1. Moments

$$\mathbb{E}X^{2n+1} = 0, \forall n \in \mathbb{N}, \mathbb{E}X^{2n} = (2n)!/(n!2^n), \forall n \in \mathbb{N}^* \quad (2.1)$$

(en particulier  $\mathbb{E}X^2 = 1$ ,  $\mathbb{E}X^4 = 3$ ) et plus généralement

$$\mathbb{E}|X|^\alpha = 2^{\alpha/2} \Gamma\{(\alpha + 1)/2\} / (\pi)^{1/2} = 2^{-\alpha/2} \Gamma(\alpha + 1) / \Gamma\{(\alpha)/2 + 1\}, \forall \alpha > -1$$

On rappelle les formules suivantes :

$$\begin{aligned} \Gamma(\alpha) &= \int_0^\infty t^{\alpha-1} \exp -tdt \\ \Gamma(n + 1) &= n! \\ \Gamma(1/2)\Gamma(\alpha) &= \Gamma(\alpha/2)\Gamma\{(\alpha + 1)/2\}2^{\alpha-1} \end{aligned}$$

2. Fonction Caractéristique - Transformée de Laplace

$$\mathbb{E} \exp -pX = \exp\{p^2/2\}, \forall p \in \mathbb{C}$$

En particulier  $\mathbb{E} \exp i\omega X = \exp\{-\omega^2/2\}, \forall \omega \in \mathbb{R}$

3. Fonction de répartition

$$F(x) = \int_{-\infty}^x \varphi(u)du, \Phi(x) = 1 - F(x)$$

Le lemme suivant donne un encadrement pour  $\Phi$  :

**Lemme 2**

$$\frac{x^2}{1+x^2} \frac{\exp \frac{-x^2}{2}}{x\sqrt{2\pi}} \leq \Phi(x) \leq \left\{ \frac{\exp \frac{-x^2}{2}}{x\sqrt{2\pi}} \right\} \wedge \left\{ \frac{\exp \frac{-x^2}{2}}{2} \right\} \quad \forall x > 0 \quad (2.2)$$

Démonstration

1. Voici d'abord une méthode que l'on peut employer de façon générale pour majorer  $P(X > x)$  : quand  $X$  a un moment exponentiel. On a par l'inégalité de Markov

$$\forall t > 0, \quad P(X > x) = P(\exp tX > \exp tx) \leq \exp -tx \mathbb{E}(\exp tX) = \exp -(tx - t^2/2).$$

On optimise cette inégalité en  $t$  en prenant  $t = x$ . D'où  $P(X > x) \leq \exp -x^2/2$ .

2. Voici maintenant une procédure plus spécifique :

$$\Phi(x) = \int_x^\infty \varphi(u) du = \int_0^\infty \exp \frac{-x^2}{2} \exp -xv \exp \frac{-v^2}{2} \frac{dv}{\sqrt{2\pi}}$$

en utilisant le changement de variable  $u = v + x$ . Maintenant, en majorant tour à tour  $\exp -xv$  puis,  $\exp \frac{-v^2}{2}$  par 1, puis en intégrant on obtient les majorations par  $\frac{\exp \frac{-x^2}{2}}{x\sqrt{2\pi}}$  puis  $\frac{\exp \frac{-x^2}{2}}{2}$ .

- 3.

$$\begin{aligned} \Phi(x) &\geq \int_x^\infty \frac{x^2}{u^2} \exp \frac{-u^2}{2} \frac{du}{\sqrt{2\pi}} \\ &= \frac{x^2}{\sqrt{2\pi}} \int_x^\infty d\left(-\frac{1}{u}\right) \exp \frac{-u^2}{2} du = \frac{x^2}{\sqrt{2\pi}} \left( \frac{1}{x} \exp \frac{-x^2}{2} - \int_x^\infty \exp \frac{-u^2}{2} du \right) \end{aligned}$$

On a utilisé une intégration par partie. On en déduit :  $\Phi(x) \geq \frac{x}{\sqrt{2\pi}} \exp \frac{-x^2}{2} - x^2 \Phi(x)$ .

Conséquences :

- Quand  $x$  tend vers l'infini,  $\Phi(x)$  est équivalent à  $\frac{\exp \frac{-x^2}{2}}{x\sqrt{2\pi}}$ .
- Pour  $x \geq 1$ , on a :

$$\frac{\exp \frac{-x^2}{2}}{2x\sqrt{2\pi}} \leq \Phi(x) \leq \frac{\exp \frac{-x^2}{2}}{x\sqrt{2\pi}}.$$

Donnons maintenant quelques valeurs utiles pour  $\Phi$ .

$x =$	0.67	1	1.96	2	3
$\Phi(x) =$	0.25	0.159	0.025	0.022	0.0015

**Définition 11**  $Y$  est une variable gaussienne réelle si il existe  $X$  gaussienne centrée, réduite,  $m \in \mathbb{R}$ ,  $\sigma^2 \in \mathbb{R}_+$  tels que

$$Y = \sigma X + m$$

*Remarques :*

1.  $\mathbb{E}Y = m$ ,  $\text{Var}Y = \sigma^2$ ,  $\mathbb{E} \exp i\omega Y = \exp i\omega m - \frac{\sigma^2 \omega^2}{2}$ ,  $\mathbb{E} \exp -pY = \exp -pm + \frac{\sigma^2 p^2}{2}$

2. On voit sur la fonction caractéristique qu'une variable gaussienne est déterminée par sa moyenne et sa variance. On notera souvent  $Y \sim N(m, \sigma^2)$ .
3. En particulier, on voit sur la définition qu'une variable constante est une variable gaussienne de variance nulle.
4. Si on applique (2.2), on obtient que si  $Y \sim N(m, \sigma^2)$ , alors  $\forall x > 0$ ,  $P(Y - m \geq x) \leq \frac{\sigma \exp \frac{-x^2}{2\sigma^2}}{x\sqrt{2\pi}}$ , par symétrie,  $P(|Y - m| \geq x) \leq \frac{2\sigma \exp \frac{-x^2}{2\sigma^2}}{x\sqrt{2\pi}}$ . On observe donc que plus  $\sigma$  augmente, moins la mesure a tendance à être concentrée.

△

## 2.2 Vecteurs gaussiens.

Par convention, un vecteur non précisé plus avant sera un vecteur colonne. Si  $u$  est un vecteur de  $\mathbb{R}^k$  ou de  $(\mathbb{R}^k)^*$ , on notera  $u^*$  son transposé.

Si  $Y = (Y_1, \dots, Y_n)^*$  est un vecteur aléatoire de  $\mathbb{R}^n$  dont les coordonnées sont intégrables, on définit le vecteur de  $\mathbb{R}^n$   $\mathbb{E}(Y) = (\mathbb{E}(Y_1), \dots, \mathbb{E}(Y_n))^*$ . Si  $X = (X_1, \dots, X_k)^*$  est un vecteur aléatoire de  $\mathbb{R}^k$ , on définit la matrice  $p \times n$  de variance-covariance  $\Sigma_{X,Y}$  dont les éléments  $a_{i,j}$  sont donnés par  $a_{i,j} = \text{cov}(X_i, Y_j)$ , si ces quantités existent.

Si  $A$  est une matrice (déterministe)  $k \times k'$ , et si  $B$  est une matrice (déterministe)  $n \times n'$  alors  $\mathbb{E}(AX) = A\mathbb{E}(X)$  et  $\Sigma_{AX, BY} = A\Sigma_{X,Y}B^*$ .

**Définition 12**  $Y = (Y_1, \dots, Y_n)^*$  est un vecteur gaussien si et seulement si toute combinaison linéaire est une gaussienne réelle.

**Exemple important :** Si, pour  $i = 1, \dots, n$ ,  $Y_i \sim N(m_i, \sigma_i^2)$  et si les  $Y_i$  sont tous mutuellement indépendants, alors on a :

$$\begin{aligned} \mathbb{E} \exp i\omega(\sum_{j=1}^n \lambda_j Y_j) &= \exp\{i\omega(\sum_{j=1}^n \lambda_j m_j) - \frac{\omega^2 \sum_{j=1}^n \lambda_j^2 \sigma_j^2}{2}\} \\ &= \exp\{i\omega(\mathbb{E} \sum_{j=1}^n \lambda_j Y_j) - \frac{\omega^2 \text{Var}(\sum_{j=1}^n \lambda_j Y_j)}{2}\} \end{aligned}$$

Ce qui entraîne  $\sum_{j=1}^n \lambda_j Y_j \sim N(\sum_{j=1}^n \lambda_j m_j, \sum_{j=1}^n \lambda_j^2 \sigma_j^2)$ , et donc  $(Y_1, \dots, Y_n)^*$  est un vecteur gaussien. Dans le cas particulier où  $m_i = 0$ ,  $\sigma_i^2 = 1$ ,  $\forall i$ , on dit que  $(Y_1, \dots, Y_n)^*$  est un vecteur gaussien standard.

**Proposition 3** Soit  $Y = (Y_1, \dots, Y_n)^*$  un vecteur aléatoire,  $m = \mathbb{E}Y = (\mathbb{E}Y_1, \dots, \mathbb{E}Y_n)^*$ ,  $V = \text{Var}Y = (\text{Cov}(Y_i, Y_j))_{i,j \in \{1, \dots, n\}}$ , alors  $Y$  est un vecteur gaussien si et seulement si sa fonction caractéristique s'écrit pour tout  $\omega^* = (\omega_1, \dots, \omega_n)$ ,

$$\mathbb{E} \exp i\omega^* Y = \exp\{i\omega^* m - \frac{\omega^* V \omega}{2}\}$$

*Remarques :*

1. La démonstration est une conséquence immédiate de la définition et du calcul de la fonction caractéristique d'une gaussienne réelle.
2. Une conséquence de cette proposition est que son espérance et sa variance déterminent un vecteur gaussien, comme dans le cas d'une variable gaussienne réelle. On notera encore  $Y \sim N(m, V)$ .

3. Si  $Y$  est un vecteur gaussien dont la matrice de covariance se sépare par blocs matriciels sous la forme suivante :

$$V = \begin{pmatrix} D_{1,i_1;1,i_1} & O_{1,i_1;i_1,i_2} & O_{1,i_1;i_2,i_3} \\ O_{i_1,i_2;1,i_1} & D_{i_1,i_2;i_1,i_2} & O_{i_1,i_2;i_2,i_3} \\ O_{i_2,i_3;1,i_1} & O_{i_2,i_3;i_1,i_2} & D_{i_2,i_3;i_2,i_3} \end{pmatrix}$$

avec  $I_{i_j,i_k;i_l,i_m}$  matrices nulles, alors les vecteurs  $(Y_1, \dots, Y_{i_1}), (Y_{i_1+1}, \dots, Y_{i_2}), (Y_{i_2+1}, \dots, Y_n)$  sont des vecteurs gaussiens, indépendants. Évidemment, la réciproque est vraie. La démonstration est immédiate (utiliser la fonction caractéristique).

4. Si  $Y \sim N(m, V)$ ,  $A$  est une matrice déterministe ( $k \times n$ ),  $b \in \mathbb{R}^k$ , alors  $Z = AY + b$  est un vecteur gaussien  $N(Am + b, AVA^*)$ . En particulier, si  $Y$  est standard ( $V = I_n$ ,  $m = 0$  où  $I_n$  est la matrice identité de  $\mathbb{R}^n$ ), alors  $Z \sim N(b, AA^*)$ .
5. Réciproquement, toute matrice de covariance est symétrique, positive, donc elle s'écrit sous la forme :

$$V = MDM^*$$

où  $M$  est une matrice orthogonale et  $D$  est une matrice diagonale, non négative. On peut donc prendre la racine de  $D = D^{1/2}$ , puis celle de  $V$  sous la forme  $V = MD^{1/2}M^* = A$ . On remarque que  $A = A^*$ ,  $AA^* = V$ . On en conclut que, suivant la construction précédente, tout vecteur gaussien s'écrit en loi sous la forme  $Y = AX + m$  où  $X$  est un vecteur gaussien standard. On voit qu'ici  $X$  est de taille  $n$  et la matrice  $A$  est  $n \times n$ . En fait, on peut préciser cette formule par la proposition suivante, qui reprend les notations ci-dessus.

△

**Proposition 4** Soit  $Y \sim N(m, V)$  un vecteur gaussien.  $V = MDM^*$  où  $M$  est une matrice orthogonale et  $D$  est une matrice diagonale, dont les coefficients diagonaux sont notés  $r_i^2$ . On suppose  $r_i^2 > 0, \forall i = 1, \dots, k$ ,  $r_i^2 = 0, \forall i \geq k + 1$ , alors, il existe  $Z \sim N(0, I_k)$ , telle que

$$Y = m + \sum_{i=1}^k r_i \vec{v}_i Z_i = m + BZ.$$

où la matrice  $B = (r_1 \vec{v}_1, \dots, r_k \vec{v}_k)$  et les  $\vec{v}_i$  sont les  $k$ -premiers vecteurs colonnes de la matrice  $M$

*Remarques :*

1. Les  $\vec{v}_i$  forment un système orthonormé.
2. La différence avec la représentation précédente ( $Y = m + AX$ ), est tout d'abord que l'écriture de la proposition est vraie presque sûrement et non plus en loi. De plus  $X$  est un vecteur gaussien standard de  $\mathbb{R}^n$  (alors que  $Z$  est un vecteur gaussien standard de  $\mathbb{R}^k$ ),  $A$  est une matrice  $n \times n$  de rang  $k$  (et donc non injective si  $k < n$ ), alors que  $B$  est ( $n \times k$ ) et injective.
3. Une conséquence de cette proposition est que la loi de  $Y$  est portée par le sous espace affine  $\{m + \sum_{i=1}^k \lambda_i \vec{v}_i, \lambda_i \in \mathbb{R}\}$  qui est de dimension  $k$ . En conséquence, si  $k < n$ , la loi de  $Y$  n'est pas absolument continue par rapport à la mesure de Lebesgue sur  $\mathbb{R}^n$ . △

**Démonstration de la Proposition :**

Posons  $\tilde{Y} = Y - m$ ,  $\tilde{Y} \sim N(0, V)$ . Soit  $W = M^* \tilde{Y}$ ,  $W \sim N(0, M^* M D M^* M) = N(0, D)$ . Donc, en fait  $W_{k+1} = \dots = W_n = 0$ . Posons  $Z_i = \frac{W_i}{r_i}$  pour  $i = 1, \dots, k$ . On a  $Z = (Z_1, \dots, Z_k)^* \sim N(0, I_k)$ . La proposition en découle. ■

On peut facilement calculer la densité de la loi de  $Y$  par rapport à la mesure de Lebesgue, quand  $V$  est inversible.

$$\phi_Y(y_1, \dots, y_n) = \frac{1}{(2\pi)^{n/2} \sqrt{\det V}} \exp\left\{-\frac{1}{2}(y - m)^* V^{-1}(y - m)\right\}$$

La démonstration se fait facilement en posant :  $y = m + Bz$ ,  $B = MD^{1/2}$ , on écrit la densité de  $z$  :  $\frac{1}{(2\pi)^{n/2}} \exp\{-\frac{1}{2}z^* z\}$ , puis on applique la formule de changement de variable.

**2.3 Normes de vecteurs gaussiens, lois  $\Gamma$** **2.3.1 Lois Gamma et quelques lois associées**

**Définition 13** On rappelle qu'une variable suit une loi  $\Gamma(p, \lambda)$ , si sa densité sur  $\mathbb{R}_+$  est :

$$1_{[0, +\infty[}(x) x^{p-1} \frac{\lambda^p}{\Gamma(p)} \exp -\lambda x$$

**Propriétés** (Démonstrations laissées en exercice)

1. La transformée de Laplace d'une loi  $\Gamma(p, \lambda)$  est définie pour  $t > -\lambda$  et vaut :

$$\mathbb{E}(\exp -tX) = \left(\frac{\lambda}{\lambda + t}\right)^p$$

2. Soit  $X \sim \Gamma(p, \lambda)$ , alors si  $a > 0$ ,  $aX \sim \Gamma(p, \lambda/a)$ .
3. Soit  $Y \sim \Gamma(q, \lambda)$ , indépendant de  $X$ . Il est clair en calculant la transformée de Laplace que  $X + Y \sim \Gamma(p + q, \lambda)$ . Il est facile de montrer que  $(X + Y, \frac{X}{X+Y})$  est un couple de variables indépendantes, ainsi d'ailleurs que le couple  $(X + Y, \frac{X}{Y})$  et que les lois respectives de  $\frac{X}{X+Y}$  et  $\frac{X}{Y}$  ne dépendent pas de  $\lambda$ . On les appellent lois Beta de première et seconde espèce.
4. La loi de  $\frac{X}{X+Y}$  (que l'on notera  $B^1(p, q)$ ,) est portée par  $[0, 1]$  et sa densité par rapport à la mesure de Lebesgue est donnée par  $\frac{x^{p-1}(1-x)^{q-1}}{B(p, q)}$ , ou  $B(p, q) = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)}$ .
5. La loi de  $\frac{X}{Y}$  (que l'on notera  $B^2(p, q)$ ,) est portée par  $[0, \infty]$  et sa densité par rapport à la mesure de Lebesgue est donnée par  $\frac{x^{p-1}}{(1+x)^{p+q} B(p, q)}$ .
6. Il est clair que si  $U \sim B^1(p, q)$ , alors  $\frac{U}{1-U} \sim B^2(p, q)$ , et que si  $V \sim B^2(p, q)$ , alors  $\frac{V}{1+V} \sim B^1(p, q)$ .
7. Il est clair que si  $V \sim B^2(p, q)$ , alors  $1/V \sim B^2(q, p)$ .

### 2.3.2 Quelques lois associées aux normes de vecteurs gaussiens :

**Définition 14** Si  $X$  est un vecteur gaussien standard de  $\mathbb{R}^n$ , alors  $\|X\|^2 = \sum_{i=1}^n X_i^2$  admet une loi appelée  $\chi^2(n)$ .

Calculons la densité de cette loi par rapport à la mesure de Lebesgue sur  $\mathbb{R}_+$  : On remarque que si  $X$  est une gaussienne réelle standard, alors  $X^2$  suit une  $\Gamma(1/2, 1/2)$ . On sait alors que si les  $X_i$  sont des copies indépendantes de  $X$ ,  $\sum_{i=1}^n X_i^2$  suit une  $\Gamma(n/2, 1/2)$ .

On en déduit que la transformée de Laplace d'une loi de  $\chi^2(n)$  existe pour  $t > -1/2$  et vaut :

$$\left(\frac{1}{1+2t}\right)^{n/2}$$

**Définition 15** – Si  $X$  suit une loi normale réelle standard,  $Y$  suit une loi  $\chi^2(k)$ , et que  $X$  et  $Y$  sont indépendants, on dit que  $T = \frac{X}{\sqrt{Y/k}}$  suit une loi de Student, notée  $T(k)$ .

– Si  $p$  et  $q$  sont des entiers, si  $X$  suit une loi de  $\chi^2(p)$  est indépendante de  $Y$  qui suit une loi de  $\chi^2(q)$ , alors  $F = \frac{X/p}{Y/q}$  suit une loi de Fisher-Snedecor à  $p$  et  $q$  degrés de liberté, notée  $F(p, q)$ .

*Exercices :*

- (a) Montrer qu'une loi  $T(n)$  admet une densité sur  $\mathbb{R}$  égale à :

$$t_n(x) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})\left(\frac{x^2}{n} + 1\right)^{\frac{n+1}{2}}}$$

Remarquer en particulier que cette loi est symétrique par rapport à l'origine.

- Montrer que  $t_n(x)$  tend quand  $n$  tend vers l'infini vers  $\varphi(x)$  pour tout  $x$  de  $\mathbb{R}$ .
  - Montrer que si  $f_n$  est une suite de fonctions réelles de la variable réelle qui converge en tout point vers une fonction  $f$ , alors si  $1 = \int f_n(x)dx = \int f(x)dx$ , la convergence a aussi lieu dans  $\mathbb{L}_1$ . C'est à dire  $\int |f_n(x) - f(x)|dx \rightarrow 0$ . (Lemme de Scheffé).
  - En déduire la convergence en loi de  $T_n$  vers une gaussienne réelle standard.
- Montrer que  $F(p, q) = q/pB^2(p/2, q/2)$  et que la densité de la loi de Fisher-Snedecor à  $p$  et  $q$  degrés de libertés est donnée par  $p/q \frac{(p/qx)^{p/2-1}}{(1+(p/qx))^{p/2+q/2}B(p, q)}$ .

$\triangle$

Remarquer par ailleurs que

- Il est clair que si  $Z \sim F(p, q)$ , alors  $1/Z \sim F(q, p)$ .
- Si  $X$  suit une loi de Student  $T(n)$ , alors  $X^2$  suit une loi de Fisher  $F(1, n)$ .

Le théorème suivant est très important pour les statistiques gaussiennes comme nous le verrons plus tard.

### 2.3.3 Théorème de Student

**Théorème 2** Soit  $X_1, \dots, X_n$ , des variables indépendantes identiquement distribuées (notation *i.i.d.*) de loi commune  $N(m, \sigma^2)$ . Alors,

- $\bar{X}_n = \frac{\sum_{i=1}^n X_i}{n}$  suit une loi  $N(m, \sigma^2/n)$ .

2.  $R_n = \sum_{i=1}^n (X_i - \bar{X}_n)^2$  suit une loi  $\sigma^2 \chi^2(n-1)$ .
3.  $\bar{X}_n$  et  $R_n$  sont indépendants.
4. Si  $S_n$  désigne la variable  $\sqrt{\frac{R_n}{n-1}}$ , alors  $T_n = \frac{\sqrt{n}(\bar{X}_n - m)}{S_n}$  suit une loi de Student  $T(n-1)$ .

### Démonstrations du Théorème de Student

- 1 est évident.
- Les quantités que nous étudions sont homogènes. Par le changement de variables  $X'_i = (X_i - m)/\sigma$ , on se ramène au cas où  $m = 0$ ,  $\sigma^2 = 1$ . Nous allons donner 3 démonstrations des points 2 et 3 chacune de ces démonstrations ayant un intérêt propre pour la suite.
- 4 est une conséquence de 2 et 3 et de la définition d'une loi de Student.

1. Première démonstration : Montrons d'abord que les vecteurs de  $\mathbb{R}$  et  $\mathbb{R}^n$ ,  $\bar{X}_n$  et  $(X_1 - \bar{X}_n, \dots, X_n - \bar{X}_n)^*$  sont indépendants. Pour cela, comme le vecteur  $(\bar{X}_n, X_1 - \bar{X}_n, \dots, X_n - \bar{X}_n)$  est gaussien, il est suffisant de calculer leur covariance :

$$\mathbb{E}\bar{X}_n(X_i - \bar{X}_n) = \mathbb{E}\bar{X}_n X_i - \mathbb{E}\bar{X}_n^2 = \mathbb{E} \frac{\sum_{j=1}^n X_j X_i}{n} - \frac{1}{n} = 0.$$

Ceci démontre 3. De plus, on a la relation suivante :

$$\sum_{i=1}^n (X_i - \bar{X}_n)^2 + n\bar{X}_n^2 = \sum_{i=1}^n X_i^2 \quad (2.3)$$

Soit  $Y + Z = W$  où  $W$  suit un  $\chi^2(n)$ ,  $Z$  suit un  $\chi^2(1)$  et  $Y$  et  $Z$  sont indépendants. Si on égale les transformées de Laplace de  $Y + Z$  et  $W$ , on en déduit facilement que  $Y$  suit un  $\chi^2(n-1)$ , ce qui démontre 2.

2. Deuxième démonstration : On considère une matrice orthogonale  $M$  telle que sa première ligne est  $(\frac{1}{\sqrt{n}}, \dots, \frac{1}{\sqrt{n}})$ . Soit  $Z = MX$  où  $X = (X_1, \dots, X_n)^*$ . Puisque  $M$  est orthogonale,  $Z$  est un vecteur gaussien standard de  $\mathbb{R}^n$ , et  $Z_1 = \sqrt{n}\bar{X}_n$  est indépendant de  $(Z_2, \dots, Z_n)$ . Par ailleurs, toujours parce que  $M$  est orthogonale,

$$\|MX\|^2 = \|X\|^2 = \sum_{i=1}^n X_i^2 = (\sqrt{n}\bar{X}_n)^2 + \sum_{i=2}^n Z_i^2.$$

On en déduit que  $\sum_{i=2}^n Z_i^2 = \sum_{i=1}^n X_i^2 - (\sqrt{n}\bar{X}_n)^2 = \sum_{i=1}^n (X_i - \bar{X}_n)^2$  (en utilisant (2.3)) est indépendant de  $\bar{X}_n$  et suit un  $\chi^2(n-1)$ .

3. Troisième démonstration. Celle-ci repose sur les deux propositions suivantes.

#### Proposition 5 .

- (a) Soit  $X \sim N(\xi, \Sigma)$  un vecteur Gaussien de  $\mathbb{R}^n$ . Soit  $A_1, A_2, \dots, A_l$ , des matrices  $n_i \times n$ . Les  $l$  vecteurs Gaussiens  $A_i X$  de  $\mathbb{R}^{n_i}$  sont mutuellement indépendants si et seulement si  $\forall i \neq j, A_i \Sigma A_j^* = 0$ .
- (b) En particulier si  $P$  est une projection orthogonale et  $X \sim N(\xi, \sigma^2 I_n)$  alors  $PX$  et  $X - PX$  sont indépendantes.

#### Proposition 6 .

- (a) Si  $P$  est une matrice de projection orthogonale (i.e.  $P = P^* = P^2$ ), et si  $W \sim N(0, P)$ , alors  $\|W\|^2 \sim \chi^2(\text{rang}(P))$
- (b) Si  $P$  est une matrice de projection orthogonale, et si  $X \sim N(0, I_n)$  alors,  $\|PX\|^2 \sim \chi^2(\text{rang}(P))$ .

Le première proposition n'est qu'une paraphrase d'une remarque antérieure. Sa démonstration est élémentaire. Elle repose essentiellement sur la fonction caractéristique.

### Démonstration de la Proposition 6 :

- (a) En effet , on peut écrire, au moyen de la matrice  $R$  orthogonale,  $P = RDR^*$  où  $D$  est une matrice diagonale dont les  $d (= \text{rang}(P))$  premiers coefficients sont égaux à 1, les autres à 0.
- Soit  $Z = R^*W$ , on a  $W = RZ$ , et  $Z \sim N(0, D)$  ce qui signifie que  $n - d$  dernières composantes sont nulles, et les  $d$  premières, sont des normales centrées réduites indépendantes. On a :  $\|W\|^2 = \|Z\|^2 = \sum_{i=1}^d Z_i^2$ .
- (b) On remarque  $PX \sim N(0, P)$ .

■

Pour finir la démonstration il suffit de remarquer que la matrice  $n \times n$ ,  $Q_n$  dont tous les termes sont égaux à  $1/n$  est une matrice de projection orthogonale de rang un, sur l'espace engendré par le vecteur  $(1, \dots, 1)^*$ . Donc  $I_n - Q_n$  est une matrice de projection de rang  $n-1$ . Les vecteurs de  $\mathbb{R}^n$ ,  $(\bar{X}_n, \bar{X}_n, \dots, \bar{X}_n)^*$ ,  $(X_1 - \bar{X}_n, \dots, X_n - \bar{X}_n)^*$  sont indépendants d'après la proposition 5 et  $\sum_{i=1}^n (X_i - \bar{X}_n)^2 \sim \sigma^2 \chi^2(n-1)$  d'après la proposition 6.

■

*Remarques :*

1. Pour illustrer la proposition précédente, remarquons qu'une matrice de projection orthogonale de rang 1 de  $\mathbb{R}^n$  est toujours de la forme  $(a_{i,j})$  avec  $a_{i,j} = a_i a_j$  et  $\sum_{i=1}^n a_i^2 = 1$ . On en conclut que si  $X \sim N(0, I_n)$  et  $\sum_{i=1}^n a_i^2 = 1$ , alors

$$\sum_{i=1}^n X_i^2 - \left( \sum_{i=1}^n a_i X_i \right)^2 \sim \chi^2(n-1).$$

Ou encore, ce qui est équivalent si  $Y \sim N(0, I_n - A)$  avec  $a_{i,j} = a_i a_j$  et  $\sum_{i=1}^n a_i^2 = 1$  alors  $\sum_{i=1}^n Y_i^2 \sim \chi^2(n-1)$ .

2. Si  $\Sigma$  est la matrice de covariance de la variable aléatoire à valeur dans  $\mathbb{R}^d$  dont la loi est donnée par  $\forall 1 \leq i \leq d$ ,  $P(Y = e_i) = p_i > 0$ , où les  $e_i$  forment la base naturelle de  $\mathbb{R}^d$ , et bien sur  $\sum_{i=1}^d p_i = 1$ . Alors si  $Z \sim N(0, \Sigma)$ ,  $\Rightarrow \sum_{i=1}^d Z_i^2 / p_i \sim \chi^2(d-1)$ . Ce point nous sera utile lors de la démonstration de la convergence du test du  $\chi^2$ .

△

### 2.3.4 Loi du $\chi^2$ décentrée

**Proposition 7** Soit  $U \sim N(\xi, I_n)$ . La loi de  $\|U\|^2$  ne dépend que de  $n$  et de  $\lambda^2 = \|\xi\|^2$ . Cette loi s'appelle un loi du  $\chi'^2(n, \lambda^2)$ .

Démonstration de la Proposition : Posons  $U = \xi + X$ ;  $X \sim N(0, I_n)$ , et calculons la transformée de Laplace de  $\|U\|^2$  :

$$\mathbb{E} \exp -p\|U\|^2 = \mathbb{E} \exp -p \sum_i^n (X_i + \xi_i)^2 = \exp(-p\|\xi\|^2) \prod_{i=1}^n \mathbb{E} \exp -p(X_i^2 + 2X_i\xi_i)$$

Par ailleurs :

$$\begin{aligned} \mathbb{E} \exp -p(X_i^2 + 2X_i\xi_i) &= \int_{-\infty}^{\infty} \exp -(px^2 + 2px\xi_i) \exp -x^2/2 \frac{dx}{\sqrt{2\pi}} \\ &= \int_{-\infty}^{\infty} \exp -1/2(x^2(1+2p) + 4px\xi_i) \frac{dx}{\sqrt{2\pi}} \\ &= \exp \frac{2p^2\xi_i^2}{2p+1} \int_{-\infty}^{\infty} \exp -\frac{1+2p}{2} \left(x + \frac{2p\xi_i}{1+2p}\right)^2 \frac{dx}{\sqrt{2\pi}} \\ &= \frac{\exp \frac{2p^2\xi_i^2}{2p+1}}{\sqrt{1+2p}} \int_{-\infty}^{\infty} \left\{ \exp -\frac{1+2p}{2} \left(x + \frac{2p\xi_i}{1+2p}\right)^2 \right\} \frac{\sqrt{1+2p}}{\sqrt{2\pi}} dx \\ &= \frac{\exp \frac{2p^2\xi_i^2}{2p+1}}{\sqrt{1+2p}} \end{aligned}$$

On en déduit :

$$\mathbb{E} \exp -p\|U\|^2 = \exp(-p\|\xi\|^2) \prod_{i=1}^n \frac{\exp \frac{2p^2\xi_i^2}{2p+1}}{\sqrt{1+2p}} = \left(\frac{1}{\sqrt{1+2p}}\right)^n \exp \frac{-p\|\xi\|^2}{2p+1}. \quad (2.4)$$

*Exercice* : On peut aussi montrer, par un argument du même type que la deuxième démonstration du théorème de Student que si  $X = \mu + X_0$ ;  $X_0 \sim N(0, I_n)$ ,  $Y = \nu + Y_0$ ;  $Y_0 \sim N(0, I_n)$ , avec  $\|\mu\|^2 = \|\nu\|^2$ , alors  $\|X\|^2 \sim \|Y\|^2$ . Pour cela on introduira une matrice orthogonale  $A$  telle que  $\nu = A\mu$ , et on utilisera :  $\|AX\|^2 = \|X\|^2$ .  $\triangle$

*Remarque* : Comme on le voit dans (2.4), la transformée de Laplace d'une loi  $\chi'^2(n, \lambda^2)$  est le produit de la transformée de Laplace d'une loi  $\chi^2(n)$  et de l'expression  $\exp \frac{-p\lambda^2}{2p+1}$ . On peut alors imaginer qu'une loi  $\chi'^2(n, \lambda^2)$  est la somme de deux lois indépendantes : une loi  $\chi^2(n)$  et une loi que l'on noterait  $\chi'^2(0, \lambda^2)$  dont la transformée de Laplace serait  $\exp \frac{-p\lambda^2}{2p+1}$ .

Montrons qu'en effet  $\exp \frac{-p\lambda^2}{2p+1}$  est la transformée de Laplace d'une loi :

Soit une suite  $(X_i)_{i \in \mathbb{N}}$  de variables indépendantes  $X_0 = 0$ , et  $\forall i \geq 1 X_i \sim N(0, 1)$ . Soit  $T$  une variable indépendante des  $X_i$ , suivant une loi de Poisson  $\mathcal{P}(\lambda^2)$ . Soit maintenant  $Y_n$  une variable qui conditionnellement à  $T = k$  suit une loi  $\chi^2(n + 2k)$ . On peut écrire qu'en loi

$$Y_n = \sum_{I=0}^{i=2T+n} X_i^2.$$

On laisse au lecteur le soin de vérifier, en calculant la transformée de Laplace, que  $\forall n \geq 1 Y_n \sim \chi'^2(n, \lambda^2)$  et que la loi de  $Y_0$  est bien la loi  $\chi'^2(0, \lambda^2)$  recherchée. On pourra vérifier aussi que cette loi, portée par  $[0, \infty)$ , est la somme d'une mesure positive portée par 0 et d'une densité dont on pourra donner une expression sous forme d'une série.  $\triangle$

**Proposition 8 .**

1. Si  $P$  est une matrice de projection (i.e.  $P = P^* = P^2$ ), et si  $W \sim N(\xi, P)$ , avec  $P(\xi) = \xi$ , alors  $\|W\|^2 \sim \chi'^2(\text{rang}(P), \|\xi\|^2)$
2. Si  $P$  est une matrice de projection (i.e.  $P = P^* = P^2$ ), et si  $X \sim N(\xi, I_n)$  alors,  $\|PX\|^2 \sim \chi'^2(\text{rang}(P), \|P(\xi)\|^2)$ .

**Démonstration de la Proposition :** La démonstration suit celle de la proposition 6

1. En effet , on peut écrire, au moyen de la matrice  $R$  orthogonale,  $P = RDR^*$  où  $D$  est une matrice diagonale dont les  $d = (\text{rang}(P))$  premiers coefficients sont égaux à 1, les autres à 0. Soit  $Z = R^*W$ . On a  $W = RZ$ , et  $Z \sim N(\eta, D)$ ,  $R^*\xi = \eta$ . Comme  $\xi = RDR^*\xi$ , on a  $\eta = D\eta$ .

Donc les  $n - d$  dernières composantes de  $Z$  sont nulles, et les  $d$  premières, suivent des lois normales  $N(\eta_i, 1)$  indépendantes. De plus  $\sum_{i=1}^n \xi_i^2 = \sum_{i=1}^d \eta_i^2$ .

Comme  $\|W\|^2 = \|Z\|^2$ ,  $\|W\|^2 \sim \chi'^2(d, \|\xi\|^2)$ .

2. On remarque  $PX \sim N(P\xi, P)$ .

■

### 2.3.5 Theoreme de COCHRAN

**Théorème 3** Soit  $X \sim N(\xi, I_n)$

1. Soit  $P_1, P_2, \dots, P_k$ ,  $k$  matrices  $n \times n$  autoadjointes, vérifiant

$$I_n = \sum_{i=1}^k P_i, \quad \text{et} \quad \sum_{i=1}^k \text{rang} P_i \leq n.$$

Alors les matrices  $P_i$  sont des projecteurs ( $P_i^2 = P_i$ ) et les variables  $P_i X$  sont des variables mutuellement indépendantes de loi  $N(P_i \xi, P_i)$ .

2. Soit  $Q_1, Q_2, \dots, Q_k$ ,  $k$  formes quadratiques sur  $\mathbf{R}^n$  vérifiant :

$$\forall x \in \mathbf{R}^n, \quad \|x\|^2 = \sum_{i=1}^k Q_i(x) \quad \text{et} \quad \sum_{i=1}^k \text{rang} Q_i \leq n.$$

Alors les variables  $Q_i(X)$  sont mutuellement indépendantes de loi  $\chi'^2(Q_i(\xi), \text{rang} Q_i)$ .

**Démonstration du Théorème :** La démonstration repose sur un lemme de pure algèbre linéaire :

**Lemme 3** Soit  $P_1, P_2, \dots, P_k$ ,  $k$  matrices  $n \times n$ , vérifiant

$$I_n = \sum_{i=1}^k P_i, \quad \text{et} \quad P_i = P_i^*$$

On a alors l'équivalence entre :

1.  $\sum_{i=1}^k \text{rang} P_i \leq n$ .
2.  $\forall i \neq j \quad P_i P_j = 0$

$$3. \forall i \quad P_i^2 = P_i$$

**Preuve du Lemme :** Remarquons que, dans ce contexte, 1 signifie que  $\sum_{i=1}^k \text{rang } P_i = n$  et :  $\forall x \in \mathbf{R}^n$ ,  $x$  s'écrit de manière unique sous la forme  $\sum_{i=1}^k u_i$ ;  $u_i \in P_i(\mathbf{R}^n)$ .

$$1. 2 \Rightarrow 3 \quad P_i = P_i(\sum_j P_j) = \sum_j P_i P_j = P_i^2$$

2. 3  $\Rightarrow$  2 On a

$$\forall x \in \mathbf{R}^n, \quad \|x\|^2 = \langle x, x \rangle = \langle x, \sum_j P_j x \rangle = \langle x, \sum_j P_j^2 x \rangle = \sum_j \|P_j x\|^2.$$

Appliquons cette relation à  $P_i x$  :

$$\forall x \in \mathbf{R}^n, \quad \|P_i x\|^2 = \sum_j \|P_j P_i x\|^2 = \|P_i x\|^2 + \sum_{j \neq i} \|P_j P_i x\|^2.$$

Donc  $j \neq i \Rightarrow P_j P_i = 0$

3. 3&2  $\Rightarrow$  1 Soit  $x = \sum_i P_i y_i$ . On a donc :

$$P_j x = \sum_i P_j P_i y_i = P_j^2 y_j = P_j y_j.$$

D'où l'écriture unique  $x = \sum_i P_i x$ .

4. 1  $\Rightarrow$  3&2  $P_j = (\sum_i P_i) P_j = \sum_i P_i P_j$ . On en déduit ;

$$\forall x \in \mathbf{R}^n, \quad P_j(x - P_j x) = \sum_{i \neq j} P_i P_j x.$$

L'unicité de la représentation implique le resultat.

### Démonstration du Théorème, (fin)

1. C'est une conséquence de la proposition 5.

2. Soit  $P_j = P_j^*$  la matrice définissant la forme quadratique  $Q_j : \forall x \in \mathbf{R}^n \quad Q_j(x) = x^* P_j x$ .

Par polarisation de la relation  $\forall x \in \mathbf{R}^n, \quad \|x\|^2 = \sum_{i=1}^k Q_i(x)$ , on obtient :

$$\forall x, y \in \mathbf{R}^n, \quad \langle x, y \rangle = \sum_j \langle x, P_j y \rangle$$

ce qui implique  $I_n = \sum_j P_j$ . Le point 2 du théorème est donc une conséquence du point 1 et de la proposition 8.

■



## Chapitre 3

# Statistique des variables gaussiennes : Introduction

Nous allons utiliser les résultats du chapitre précédent pour introduire des procédures statistiques dans le cadre d'un échantillon gaussien. Notre point de vue est de développer ici une vision rapide et intuitive. En particulier, les définitions formelles seront données dans le chapitre suivant.

### 3.1 Estimation de la moyenne d'un échantillon gaussien

En statistique, on se pose généralement 2 types de problèmes : Estimation ou test. Dans ce chapitre, nous n'envisagerons que le premier.

Prenons l'exemple suivant de l'estimation de la moyenne d'un  $n$ - échantillon gaussien :

Supposons que l'on observe  $n$  données  $x_1, \dots, x_n$  qui chacune représente une mesure d'une quantité physique  $\mu$ , inconnue que l'on cherche à estimer. Chacune de ces données  $x_i$  est entachée d'une erreur due à la mesure.

Nous supposons ici que les erreurs  $\varepsilon_i$  sont indépendantes, identiquement distribuées de loi  $N(0, \sigma^2)$ , de sorte que  $X_1, \dots, X_n$  sont i.i.d.  $N(\mu, \sigma^2)$ . Notons  $P_{\mu, \sigma^2}^n$  la loi de l'échantillon  $X_1, \dots, X_n$ .

Dans ces conditions, un **estimateur** naturel de  $\mu$  est

$$\bar{X}_n = \frac{\sum_{i=1}^n X_i}{n},$$

(ce qui calculé sur nos observations, donne la valeur  $\bar{x}_n = \frac{\sum_{i=1}^n x_i}{n}$ .)

Outre le fait que cet estimateur est 'naturel', on peut remarquer qu'il est **convergent**. En effet, la loi faible des grands nombres nous dit que pour tous  $\varepsilon > 0$ ,  $\mu$ ,  $\sigma^2$ ,

$$\lim_{n \rightarrow \infty} P_{\mu, \sigma^2}^n(|\bar{X}_n - \mu| \geq \varepsilon) = 0$$

## 3.2 Intervalles de confiance pour l'estimation de la moyenne.

Mais la donnée d'une valeur d'estimation n'est pas suffisamment informative si on ne lui adjoint pas une idée de la précision de cette estimation. Pour cela nous allons maintenant sur cet exemple construire un **intervalle de confiance** : Il est facile de vérifier que  $\bar{X}_n$  suit une loi  $N(\mu, \sigma^2/n)$ . Ceci nous permet de calculer notre probabilité d'erreur : Considérons deux cas : le premier cas où  $\sigma^2$  est connu (on pourra donc l'utiliser dans la recherche d'un intervalle de confiance), le deuxième cas où  $\sigma^2$  n'est pas connu.

### 3.2.1 $\sigma^2$ connu.

Si  $\sigma$  est connu, on a  $P_{\mu, \sigma^2}^n \left( \frac{\sqrt{n}|\bar{X}_n - \mu|}{\sigma} \geq t \right) = 2\Phi(t)$ . En particulier, si on prend  $t=1.96$ , on trouve

$$\forall \mu \in \mathbf{R}, P_{\mu, \sigma^2}^n \left( \bar{X}_n - \frac{1.96\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}_n + \frac{1.96\sigma}{\sqrt{n}} \right) = 0.95.$$

On dit alors que l'intervalle aléatoire

$$\left[ \bar{X}_n - \frac{1.96\sigma}{\sqrt{n}}, \bar{X}_n + \frac{1.96\sigma}{\sqrt{n}} \right]$$

est un **intervalle de confiance, au niveau d'erreur 0.05**.

On sous entend ainsi que la valeur cherchée  $\mu$  n'a que 0.05 de chance de ne pas être dans cet intervalle. On donne alors à l'utilisateur la fourchette suivante calculée sur ses propres données  $[\bar{x}_n - \frac{1.96\sigma}{\sqrt{n}}, \bar{x}_n + \frac{1.96\sigma}{\sqrt{n}}]$ . La longueur de cet intervalle est  $\frac{1.96\sigma}{\sqrt{n}}$ .

On voit donc que la précision augmente quand on augmente le nombre d'observations et diminue quand  $\sigma$  augmente.

### 3.2.2 $\sigma^2$ inconnu.

Si  $\sigma$  est inconnu, la méthode précédente ne s'applique plus puisque l'intervalle précédent était exprimé en fonction de  $\sigma$ . Néanmoins, le théorème de Student va encore nous permettre de construire un intervalle de confiance.

En effet, on a vu que  $T_n = \frac{\sqrt{n}(\bar{X}_n - m)}{S_n}$  suit une loi  $T(n-1)$ . Notons

$$\Phi_k(x) = P(Z_k > x)$$

si  $Z_k$  a la loi  $T(k)$ .

Dans les tables, qui figurent à la fin de ce livre où dans les logiciels, on peut trouver  $z_{n-1}$  tel que  $2\Phi_{n-1}(z_{n-1}) = 0.05$ .

Comme  $P_{\mu, \sigma^2}^n \left( \frac{\sqrt{n}|\bar{X}_n - \mu|}{S_n} \geq z_{n-1} \right) = 2\Phi_{n-1}(z_{n-1})$ , on aura,

$$\forall \mu \in \mathbf{R}, \sigma^2 \in \mathbf{R}_+, P_{\mu, \sigma^2}^n \left( \bar{X}_n - \frac{z_{n-1}S_n}{\sqrt{n}} \leq \mu \leq \bar{X}_n + \frac{z_{n-1}S_n}{\sqrt{n}} \right) = 0.95.$$

et l'intervalle aléatoire

$$\left[ \bar{X}_n - \frac{z_{n-1}S_n}{\sqrt{n}}, \bar{X}_n + \frac{z_{n-1}S_n}{\sqrt{n}} \right]$$

est un **intervalle de confiance**, au niveau d'erreur **0.05**.

*Exercice :* Calculer dans le cas  $\sigma$  connu, quel est le nombre minimum d'observations nécessaire pour obtenir une fourchette au niveau d'erreur 0.05 dont la longueur ne dépasse pas une quantité fixée  $\delta$ . (réponse :  $n \geq (\frac{2 \cdot (1.96)\sigma}{\delta})^2$ .)  $\triangle$

### 3.3 Estimation de la variance d'un échantillon gaussien

On peut aussi dans le cas où  $\sigma^2$  est inconnu vouloir l'estimer et construire des intervalles de confiance. Comme précédemment nous distinguerons les cas  $\mu$  connu et inconnu.

#### 3.3.1 $\mu$ connu.

On remarque que

$$R'_n{}^2 = \sum_{i=1}^n (X_i - \mu)^2$$

suit sous  $P_{\mu, \sigma^2}^n$  une loi de  $\sigma^2 \chi^2(n)$ . Si on introduit

$$\hat{\sigma}_n^2 = \frac{R'_n{}^2}{n},$$

il est facile de vérifier que

$$\forall \mu \in \mathbf{R}, \sigma^2 \in \mathbf{R}_+, \mathbf{E}_{\mu, \sigma^2}^n \hat{\sigma}_n^2 = \sigma^2$$

Une telle propriété nous dit que la statistique  $\hat{\sigma}_n^2$  est un 'estimateur' de  $\sigma^2$  qui a la propriété d'être 'en moyenne' égal à la quantité cherchée. On dit qu'il est **sans biais**.

Cherchons maintenant un intervalle de confiance pour  $\sigma^2$ . Comme précédemment, notons

$$\Psi_k(x) = P(Z_k > x)$$

si  $Z_k$  a la loi  $\chi^2(k)$ .

Dans les tables, qui figurent à la fin de ce livre ou dans les logiciels, on peut trouver  $z_n$  et  $z'_n$  tels que  $\Psi_n(z_n) = 0.025$  et  $\Psi_n(z'_n) = 0.975$ . On a alors  $P_{\mu, \sigma^2}^n \left( \frac{n\hat{\sigma}_n^2}{\sigma^2} \in [z'_n, z_n] \right) = 0.95$ . De sorte que :

$$\forall \mu \in \mathbf{R}, \sigma^2 \in \mathbf{R}_+, P_{\mu, \sigma^2}^n \left( \frac{n\hat{\sigma}_n^2}{z_n} \leq \sigma^2 \leq \frac{n\hat{\sigma}_n^2}{z'_n} \right) = 0.95.$$

Donc  $\left[ \frac{n\hat{\sigma}_n^2}{z_n}, \frac{n\hat{\sigma}_n^2}{z'_n} \right]$  est un intervalle de confiance au niveau d'erreur 0.05.

#### 3.3.2 $\mu$ inconnu.

La démarche est presque semblable, sauf que nous allons maintenant utiliser de théorème de Student. En utilisant ce théorème, on démontre en effet que :

$$R_n^2 = \sum_{i=1}^n (X_i - \bar{X}_n)^2$$

suit sous  $P_{\mu, \sigma^2}^n$  une loi de  $\sigma^2 \chi^2(n-1)$ . Pour

$$S_n^2 = \frac{R_n^2}{n-1},$$

il est facile de vérifier que

$$\forall \mu \in \mathbf{R}, \sigma^2 \in \mathbf{R}_+, \mathbf{E}_{\mu, \sigma^2}^n S_n^2 = \sigma^2$$

c'est à dire que l'estimateur  $S_n^2$  est sans biais.

Cherchons maintenant un intervalle de confiance. Comme précédemment, considérons  $z_{n-1}$  et  $z'_{n-1}$  tels que  $\Psi_{n-1}(z_{n-1}) = 0.025$  et  $\Psi_{n-1}(z'_{n-1}) = 0.975$ .

On a alors  $P_{\mu, \sigma^2}^n \left( \frac{(n-1)S_n^2}{\sigma^2} \in [z'_{n-1}, z_{n-1}] \right) = 0.95$ . De sorte que :

$$\forall \mu \in \mathbf{R}, \sigma^2 \in \mathbf{R}_+, P_{\mu, \sigma^2}^n \left( \frac{(n-1)S_n^2}{z_{n-1}} \leq \sigma^2 \leq \frac{(n-1)S_n^2}{z'_{n-1}} \right) = 0.95.$$

Donc  $\left[ \frac{(n-1)S_n^2}{z_{n-1}}, \frac{(n-1)S_n^2}{z'_{n-1}} \right]$  est un intervalle de confiance au niveau d'erreur 0.05.

## Chapitre 4

# METHODES DE SUBSTITUTION, CONVERGENCES.

Au cours des chapitres qui suivent, nous allons essentiellement décrire des méthodes classiques d'estimation. Plus précisément, nous étudierons principalement les méthodes de **substitutions**, les méthodes de **contrastés** et les méthodes **bayésiennes**. Les deux dernières méthodes sont, sur le plan conceptuel, comme sur le plan algorithmique plus compliquées à mettre en oeuvre, plus précises aussi dans une grande majorité des cas.

La méthode de substitution que nous allons présenter ici est assez simple à comprendre et à mettre en oeuvre dans la pratique. Il est de plus souvent assez facile d'évaluer ses performances et son domaine de fiabilité.

Elle nous permettra en outre de définir les propriétés classiques de **convergence** et de **convergence en loi**, qui sont fort utiles en statistique. Elle nous permettra aussi de définir les notions fondamentales **d'intervalles de confiance exacts et asymptotiques** et de construire de tels intervalles sous des conditions assez larges

### 4.1 DÉFINITIONS

#### 4.1.1 Estimateurs

Nous disposons d'une expérience

$$\mathcal{E} = (X, \mathcal{X}, \mathcal{A}, P_\theta, \theta \in \Theta),$$

Notre propos, dans ce chapitre va être d'estimer une quantité  $q(\theta)$  où  $q$  est une application de  $\Theta$  dans un espace  $E$ , muni d'une tribu  $\mathcal{B}$  et d'une métrique  $d$ .

**Définition 16** *Dans le cadre précédemment énoncé, un estimateur de la quantité  $q(\theta)$  est une variable aléatoire de la forme*

$$T = \phi \circ X$$

où  $\phi$  est une fonction mesurable de  $\mathcal{X}, \mathcal{A}$  dans  $E, \mathcal{B}$ .

On voit immédiatement que la définition n'est pas contraignante. En revanche, nous apprendrons rapidement à mesurer les performances des différents estimateurs, ce qui exclura les estimateurs "fantaisistes".

### 4.1.2 Estimateurs de substitution.

Considérons le **modèle d'échantillonnage** associé à l'expérience  $\mathcal{E}$ ,

$$\mathcal{E}^n = (X^n, \mathcal{X}^n, \mathcal{A}_n, P_\theta^n, \theta \in \Theta)$$

$P_\theta^n$  est la probabilité produit de  $n$  copies indépendantes de la loi  $P_\theta$ , notée aussi  $P_\theta^{\otimes n}$ .

Nous observons donc  $X^n = (X_1, \dots, X_n)$  où les  $X_i$  sont i.i.d..

**Définition 17** *Supposons qu'il existe :*

1.  $f_1, \dots, f_r$ ,  
 $r$  fonctions mesurables de  $\mathcal{X}, \mathcal{A}$  dans  $\mathbf{R}, \mathcal{B}(\mathbf{R})$  telles que  
pour tout  $\theta$  dans  $\Theta$  :

$$\mathbf{E}_\theta |f_j(X)| = \int_\Omega |f_j(X)(\omega)| dP_\theta(\omega) < \infty, \forall j \in \{1, \dots, r\}. \quad (4.1)$$

2.  $g$  fonction continue de  $\mathbf{R}^r$  dans  $E$ , telle que **pour tout  $\theta$  dans  $\Theta$ ,**

$$q(\theta) = g(\mathbf{E}_\theta f_1(X), \dots, \mathbf{E}_\theta f_r(X))$$

Définissons pour  $j \in \{1, \dots, r\}$ , la variable aléatoire,

$$\hat{f}_j = \frac{1}{n} \sum_{i=1}^n f_j(X_i)$$

On appelle **estimateur de substitution** de la quantité  $q(\theta)$ ,

$$T_n = g(\hat{f}_1, \dots, \hat{f}_r).$$

## 4.2 Exemples d'estimateurs de substitution.

Dans le cas où  $r = 1$ , en prenant pour fonction  $g$ , l'identité, on voit que  $\hat{f}_1$  est lui même un estimateur de substitution, de la quantité  $E_\theta f_1(X)$ .

Suivant la forme des fonctions  $f_j$ , les estimateurs peuvent porter des noms différents. Etudions les cas particuliers suivants.

### 4.2.1 Substitution des fréquences.

1. Si les  $X_i$  sont des variables aléatoires réelles,  $t$  un point arbitraire, fixé de  $\mathbf{R}$ , en prenant  $f_1(u) = I\{] \infty, t\}(u)$ , on voit que l'hypothèse (4.1) est vérifiée quelque soit l'ensemble  $\Theta$  dont on est parti dans le modèle, puisque  $f_1$  est bornée. On obtient donc que

$$\hat{F}_n(t) = \frac{1}{n} \sum_{i=1}^n I\{X_i \leq t\}$$

est un estimateur de substitution de la quantité

$$F_\theta(t) = E_\theta I\{X \leq t\} = P_\theta(X \leq t).$$

La fonction aléatoire  $\hat{F}_n$  ainsi définie pour tout  $t$  dans  $\mathbf{R}$ , est appelée **fonction de répartition empirique**. Elle estime point par point la fonction de répartition théorique  $F_\theta$ , quelque soit l'ensemble  $\Theta$  dont on est parti dans le modèle.

2. Soit  $A_1, \dots, A_r$  une partition mesurable de  $\mathcal{X}, \mathcal{A}$ .

$$\hat{p}_j = \frac{1}{n} \sum_{i=1}^n I\{X_i \in A_j\}$$

représente la fréquence avec laquelle l'échantillon  $(X_1, \dots, X_n)$  visite l'ensemble  $A_j$ . Comme dans l'exemple précédent, sans aucune hypothèse sur le modèle, c'est un estimateur par substitution de la quantité  $P_\theta(A_j)$ .

Si par ailleurs, nous savons par hypothèse qu'une quantité d'intérêt  $q(\theta)$  s'écrit comme

$$q(\theta) = g(P_\theta(A_1), \dots, P_\theta(A_r)),$$

avec  $g$ , fonction continue, il est clair que

$$T_n = g(\hat{p}_1, \dots, \hat{p}_r),$$

est un estimateur par substitution de  $q(\theta)$ .

3. Donnons un exemple pratique de ce dernier cas particulier : Le modèle de **Hardy-Weinberg**.

On suppose qu'un gène comporte 2 allèles A et a. L'allèle A apparait dans la population avec une proportion  $\theta \in [0, 1]$ , inconnue. Les 2 allèles se combinent sous la forme AA, Aa ou aa. Ces trois combinaisons donnent lieu à des phénotypes identifiables. Par exemple AA est malade, Aa n'est pas malade mais porteur identifiable par des analyses biologiques, aa est sain.

On veut estimer la quantité

$$q(\theta) = \theta.$$

Pour cela on observe pour  $n$  individus indépendants dans la population, la classe où ils se trouvent.

On est automatiquement sous la forme de l'exemple précédent avec  $A_1 = AA$ ,  $A_2 = Aa$ ,  $A_3 = aa$ .  $\hat{p}_1$  et  $\hat{p}_3$  représentent les proportions observées d'homozygotes malades ou sains,  $\hat{p}_2$  représente la proportion observée d'hétérozygotes. Le modèle de Hardy-Weinberg consiste à faire l'hypothèse suivante,

$$\begin{aligned} P_\theta(A_1) &= \theta^2 \\ P_\theta(A_2) &= 2\theta(1 - \theta) \\ P_\theta(A_3) &= (1 - \theta)^2 \end{aligned}$$

Nous pouvons par exemple remarquer que notre quantité d'intérêt  $\theta$  s'écrit comme  $\sqrt{P_\theta(A_1)}$ . On en déduit donc que

$$T_n^1 = \sqrt{\hat{p}_1}$$

est un estimateur par substitution des fréquences de  $\theta$ .

On aurait pu aussi remarquer que  $\theta$  s'écrit aussi  $1 - \sqrt{P_\theta(A_3)}$ , ce qui nous fournit immédiatement un nouvel estimateur par substitution des fréquences :

$$T_n^2 = 1 - \sqrt{\hat{p}_3}.$$

Bien entendu,  $T_n^1$  et  $T_n^2$  ne coïncident pas en général, et cela crée une ambiguïté, puisque la méthode peut définir plusieurs estimateurs, et qu'il nous faudra ensuite choisir.

### 4.2.2 Substitution des moments.

Un autre cas particulier important de la méthodes de substitution est la méthode des moments. Elle consiste simplement à considérer le cas particulier où :

$$\begin{aligned}(\mathcal{X}, \mathcal{A}) &= (\mathbf{R}, \mathcal{B}(\mathbf{R})) \\ f_j(x) &= x^j, \forall j \in \{1, \dots, r\}.\end{aligned}$$

La condition (4.1) devient alors :

$$\forall \theta \in \Theta, \quad \mathbf{E}_\theta |X|^r = \int_{\Omega} |X(\omega)|^r dP_\theta(\omega) < \infty. \quad (4.2)$$

Donnons un exemple. Supposons que dans une expérience dont les données sont réelles, on ait la condition :

$$\forall \theta \in \Theta, \quad \mathbf{E}_\theta |X|^2 < \infty.$$

Si l'on se propose d'estimer la quantité (paramètre de centralité) :

$$\mu(\theta) = \mathbf{E}_\theta X,$$

l'estimateur des moments est :

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i.$$

Si l'on se propose d'estimer le paramètre de dispersion :

$$\sigma^2(\theta) = \mathbf{E}_\theta (X - \mathbf{E}_\theta X)^2 = \mathbf{E}_\theta X^2 - (\mathbf{E}_\theta X)^2,$$

l'estimateur des moments est :

$$S_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i)^2 - \left( \frac{1}{n} \sum_{i=1}^n X_i \right)^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2.$$

## 4.3 Convergence d'une suite d'estimateurs.

### 4.3.1 Convergence dans un cadre général.

– On se donne une suite générale d'expériences de la forme,

$$\mathcal{E}_n = (\Omega_n, \mathcal{F}_n, X^n, \mathcal{X}_n, \mathcal{A}_n, P_\theta^n, \theta \in \Theta),$$

dans laquelle tout peut varier avec  $n$  sauf l'ensemble des paramètres. Une suite de  $n$  échantillonnages est un exemple de telle suite.

- On se propose d'estimer une quantité  $g(\theta)$  appartenant à un espace  $E$  muni d'une tribu  $\mathcal{B}$  et que l'on suppose métrique (i.e. muni d'une distance  $d$ ).
- On dispose d'une suite d'estimateurs  $T_n$  **adaptée à  $\mathcal{E}_n$**  c'est à dire de la forme  $T_n = h_n(X^n)$  où  $h_n$  est une suite d'applications mesurables de  $\mathcal{X}_n, \mathcal{A}_n$  dans  $E, \mathcal{B}$ .

**Définition 18** Dans le contexte ci-dessus, on dit que la suite d'estimateurs de  $q(\theta)$ ,  $T_n$  **converge le long de la suite d'expériences**  $\mathcal{E}_n$  si

Pour tout  $\theta$  appartenant à  $\Theta$ ,  $T_n$  converge en  $P_\theta^n$ - probabilité vers  $q(\theta)$ , i.e. :

$$\forall \theta \in \Theta, \quad \forall \epsilon > 0, P_\theta^n \{d(T_n, q(\theta)) > \epsilon\} \longrightarrow 0.$$

Nous avons alors la proposition,

**Proposition 9** Sous les conditions de la définition 17, l'estimateur de substitution  $T_n$  converge le long de l'expérience échantillonnée  $\mathcal{E}^n$ .

**Démonstration de la Proposition :** Cette démonstration est une simple conséquence de la continuité de la fonction  $g$  alliée à la loi des grands nombres, qui implique que

$$\forall \theta \in \Theta, \quad \forall j \in \{1, \dots, r\}, \quad \hat{f}_j \xrightarrow{P_\theta^n - \text{prob}} \mathbf{E}_\theta f_j,$$

puisque  $\mathbf{E}_\theta |f_j| < \infty$ .

*Remarque :* Dans le cas précédent, la convergence a lieu de façon plus forte, puisque  $\hat{f}_j$  converge aussi presque sûrement. Notons toutefois que la convergence presque sûre peut ne pas avoir de sens dans une suite générale d'expérience, puisque dans ce cas les espaces  $\mathcal{X}_n$  sont arbitraires, il n'est donc pas forcément question de les emboîter et de les inclure tous dans un plus gros espace, sur lequel les  $P_\theta^n$  sont des probabilités traces.  $\triangle$

La propriété de convergence est intéressante puisqu'elle nous dit que plus on a d'observations, plus on a "de chances d'être proche de la quantité que l'on estime." En revanche, elle ne dit rien de quantifiable sur le comportement de l'estimateur et sa précision. Nous allons dans les paragraphes suivants nous efforcer de préciser cette notion.

## 4.4 Intervalles de confiance.

Supposons maintenant que la quantité à estimer  $q(\theta)$  soit réelle.

**Définition 19** Soit  $\alpha$  fixé dans  $(0, 1)$ . Soit, dans une expérience arbitraire  $\mathcal{E}$  (échantillonnée ou non),  $S = h_o X$ ,  $T = h' o X$ , 2 estimateurs de  $q(\theta)$ , on dira que  $[S, T]$  est un **intervalle de confiance au niveau  $\alpha$** , si

$$\forall \theta \in \Theta, \quad P_\theta \{q(\theta) \in [S, T]\} \geq 1 - \alpha.$$

*Remarque :* Bien entendu,  $S = -\infty$ ,  $T = \infty$  convient toujours mais n'est guère intéressant. En effet, notre intérêt pratique sera toujours de rendre  $T - S$  le plus petit possible.  $\triangle$

Dans le paragraphe suivant, nous allons proposer une construction d'intervalles de confiance dans certains cas pour les méthodes de substitution.

#### 4.4.1 Intervalles de confiance associés à l'inégalité d'Hoeffding.

Nous commençons par l'énoncé du théorème suivant :

##### **Théorème 4 Inégalité de Hoeffding**

Soit  $Y_1, \dots, Y_n$ ,  $n$  variables aléatoires indépendantes vérifiant

$$a_i \leq Y_i \leq b_i; \quad \mathbf{E}(Y_i) = 0.$$

On note  $\Delta_i = b_i - a_i$ . Alors :

$$\begin{aligned} \forall \lambda > 0; \quad P[\sum_{i=1}^n Y_i \geq \lambda] &\leq \exp - \frac{2\lambda^2}{\sum_{i=1}^n \Delta_i^2} \\ P[|\sum_{i=1}^n Y_i| \geq \lambda] &\leq 2 \exp - \frac{2\lambda^2}{\sum_{i=1}^n \Delta_i^2} \end{aligned}$$

**Preuve :**

$\forall t > 0, \forall \lambda > 0,$

$$\begin{aligned} P[\sum_{i=1}^n Y_i \geq \lambda] &= P[\exp t(\sum_{i=1}^n Y_i) \geq \exp t\lambda] \\ &\leq \exp -t\lambda \mathbf{E}[\exp t(\sum_{i=1}^n Y_i)] \\ &\leq \exp -t\lambda \prod_{i=1}^n \mathbf{E}[\exp tY_i] \\ &\leq \exp -[t\lambda - \sum_{i=1}^n \log(\mathbf{E}(\exp tY_i))] \end{aligned}$$

**Lemme 3** Si  $a \leq Y \leq b$ ,  $\mathbf{E}(Y) = 0$ , et  $\Delta = b - a$ ; alors

$$\log(\mathbf{E}(\exp tY)) \leq \frac{t^2}{8} \Delta^2$$

**Preuve du lemme 5 :** On remarquera que  $\mathbf{E}(Y) = 0$  implique  $a \leq 0$ . Par ailleurs on a évidemment :

$$Y = a \frac{b - Y}{b - a} + b \frac{Y - a}{b - a}.$$

Soit  $t > 0$  fixé. Par la convexité de la fonction  $y \rightarrow \exp ty$  il vient

$$\exp tY \leq \frac{b - Y}{b - a} \exp ta + \frac{Y - a}{b - a} \exp tb.$$

En prenant l'esperance, et comme  $\mathbf{E}(Y) = 0$ ,

$$\mathbf{E}(\exp tY) \leq \frac{b}{b - a} \exp ta - \frac{a}{b - a} \exp tb.$$

Si on pose  $0 \leq \alpha = \frac{-a}{b-a} \leq 1$ , on peut réécrire l'inégalité précédente :

$$\begin{aligned} \mathbf{E}(\exp tY) &\leq (1 - \alpha) \exp -at\Delta + \alpha \exp t(1 - \alpha)\Delta. \\ \log(\mathbf{E}(\exp tY)) &\leq -at\Delta + \log(1 - \alpha + \alpha \exp t\Delta). \end{aligned}$$

En posant  $x = t\Delta$ , on peut conclure grâce au lemme suivant. ■

**Lemme 4** Soit  $0 \leq \alpha \leq 1$ .

$$\forall x \geq 0, \quad [-\alpha x + \log(\alpha \exp x + 1 - \alpha)] \leq \frac{1}{8}x^2.$$

**Preuve du lemme 6 :** Soit  $0 \leq \alpha \leq 1$ , fixé et soit

$$f(x) = \frac{1}{8}x^2 - [-\alpha x + \log(\alpha \exp x + 1 - \alpha)].$$

Il nous faut démontrer que  $f(x) \geq 0$ . On vérifie aisément que

$$f(0) = 0; \quad f'(0) = 0; \quad f''(x) = 1/4 - \frac{(\alpha \exp x)(1 - \alpha)}{[(\alpha \exp x) + (1 - \alpha)]^2}$$

Cette dernière quantité est toujours positive, puisque  $\sup_{a>0; b>0} \frac{ab}{(a+b)^2} = 1/4$ . Il s'en suit que  $\forall x \geq 0, \quad f'(x) \geq 0, \quad f(x) \geq 0$ . ■

**Fin de la démonstration :**

On a donc

$$P\left[\sum_{i=1}^n Y_i \geq \lambda\right] \leq \exp\left[-t\lambda - \frac{t^2}{8} \sum_{i=1}^n \Delta_i^2\right]$$

En optimisant en  $t$  (c'est à dire en prenant  $t \sum_{i=1}^n \Delta_i^2 / 4 = \lambda$ ), on obtient le résultat. La deuxième inégalité s'obtient simplement en remarquant que pour  $\lambda > 0$  :

$$P\left[\left|\sum_{i=1}^n Y_i\right| \geq \lambda\right] = P\left[\sum_{i=1}^n Y_i \geq \lambda\right] + P\left[\sum_{i=1}^n -Y_i \geq \lambda\right].$$

■

Ce théorème nous permet immédiatement de déduire la proposition suivante :

**Proposition 10** Sous les conditions de la définition 17, si l'on estime la quantité

$$q(\theta) = \mathbf{E}_\theta f_1(X),$$

et que  $f_1$  est uniformément borné par la constante  $M$ , alors si  $T_n$  est l'estimateur de substitution,

$$[T_n - t_{n,\alpha}, T_n + t_{n,\alpha}], \quad t_{n,\alpha} = \sqrt{\frac{8M^2}{n} \log(2/\alpha)}$$

est un intervalle de confiance au niveau  $\alpha$ .

**Démonstration de la Proposition.**

Il suffit de remarquer que

$$\begin{aligned} P_\theta\{q(\theta) \in [T_n - t_{n,\alpha}, T_n + t_{n,\alpha}]\} &= P_\theta\{T_n - q(\theta) \in [-t_{n,\alpha}, +t_{n,\alpha}]\} \\ &= P_\theta\{|\frac{1}{n} \sum_{i=1}^n [f_1(X_i) - \mathbf{E}_\theta f_1]| \leq t_{n,\alpha}\}. \end{aligned}$$

On peut alors appliquer l'inégalité de Hoeffding avec  $Y_i = [f_1(X_i) - \mathbf{E}_\theta f_1]$ .

On a alors  $\Delta_i = 4M$ , et en posant  $\lambda = nt_{n,\alpha}$ , le résultat annoncé.

*Remarques :*

1. On remarque que la longueur de l'intervalle de confiance construit est exactement  $2t_{n,\alpha}$ , c'est à dire qu'elle est croissante en  $M$  et décroissante en fonction du nombre d'observations comme  $1/\sqrt{n}$ .
2. Si l'on reprend l'exemple qui consistait à estimer la fonction de répartition (soit, la quantité  $q(\theta) = F_\theta(x)$ ) en un point arbitraire fixé  $x$ , nous avons clairement  $M = 1$ , et on en déduit que sans hypothèses autre que le fait que l'on observe un modèle réel échantillonné, nous avons que

$$[\hat{F}_n(x) - \sqrt{\frac{8}{n} \log(2/\alpha)}, \hat{F}_n(x) + \sqrt{\frac{8}{n} \log(2/\alpha)}],$$

est un intervalle de confiance de niveau  $\alpha$ .

3. Démontrer en exercice qu'exactement les mêmes résultats peuvent être obtenus si l'on désire estimer, à la place de la fonction de répartition, la fonction caractéristique :

$$\mathbf{E}_\theta \exp\{itX\}$$

en un point  $t$  fixé.  $\triangle$

Cette méthode donne de bons résultats. Malheureusement l'hypothèse  $f_1$  uniformément borné est souvent mise en défaut, par exemple dans la cas simple de l'estimation du paramètre moyenne pour une loi réelle qui n'est pas à support compact. Le paragraphe suivant propose de régler ce problème dans le cadre moins restrictif des intervalles de confiance asymptotiques.

## 4.5 Intervalles de confiance asymptotiques.

Nous supposons toujours que la quantité à estimer  $q(\theta)$  est réelle.

**Définition 20** Soit  $\alpha$  fixé dans  $(0, 1)$ . On se donne une suite générale d'expériences de la forme,

$$\mathcal{E}_n = (\Omega_n, \mathcal{F}_n, X^n, \mathcal{X}_n, \mathcal{A}_n, P_\theta^n, \theta \in \Theta),$$

(échantillonnée ou non), soit  $S_n = h_n o X^n$ ,  $T_n = h'_n o X^n$ , 2 suites d'estimateurs de  $q(\theta)$  adaptées à  $\mathcal{E}_n$ , on dira que  $[S_n, T_n]$  est **une suite d'intervalles de confiance asymptotiquement de niveau  $\alpha$**  (notée  $ICA(\alpha)$ ), si

$$\forall \theta \in \Theta, \quad P_\theta^n \{q(\theta) \in [S_n, T_n]\} \longrightarrow 1 - \alpha.$$

*Remarque* : Il est clair qu'avoir un intervalle de confiance asymptotiquement de niveau  $\alpha$  est beaucoup moins prudent qu'avoir intervalle de confiance de niveau  $\alpha$ . En effet, on ne controle pas exactement l'erreur, on ne fait que dire qu'elle est asymptotiquement de l'ordre fixé. Mais en particulier on ne précise pas à partir de quel nombre de données l'approximation devient raisonnable. Ceci peut se faire, à l'aide par exemple de précision dans la convergence par l'intermédiaire de développements en fonction de puissances de  $1/\sqrt{n}$ , de type Berry-Esseen ou Edgeworth. (cf, par exemple Feller, Tome 2 <sup>1</sup>, Petrov <sup>2</sup>.)  $\triangle$

Nous allons mettre en place des intervalles de confiance asymptotiques pour les méthodes de substitutions sous d'assez larges hypothèses. Pour cela, il est fondamental de définir la notion suivante.

<sup>1</sup>Feller, William, An introduction to probability theory and its applications. , John Wiley & Sons Inc., New York, 1971,

<sup>2</sup>Petrov, Valentin V., Limit theorems of probability theory, The Clarendon Press Oxford University Press, New York, 1995,

## 4.6 Convergence en loi d'une suite d'estimateurs.

- Comme précédemment, on se donne une suite générale d'expériences de la forme,

$$\mathcal{E}_n = (\Omega_n, \mathcal{F}_n, X^n, \mathcal{X}_n, \mathcal{A}_n, P_\theta^n, \theta \in \Theta).$$

- On se propose d'estimer une quantité  $q(\theta)$  appartenant à  $\mathbf{R}$ .
- On dispose d'une suite d'estimateurs  $T_n$  **adaptée à  $\mathcal{E}_n$**  c'est à dire de la forme  $T_n = h_n(X^n)$  où  $h_n$  est une suite d'applications mesurables de  $\mathcal{X}_n, \mathcal{A}_n$  dans  $E, \mathcal{B}$ .

**Définition 21** *Étant donnée  $c_n$  une suite déterministe positive tendant vers l'infini, on dit que la suite d'estimateurs de  $q(\theta)$ ,  $T_n$  converge en loi le long de la suite d'expériences  $\mathcal{E}_n$  à la vitesse  $c_n$  si*

*Pour tout  $\theta$  appartenant à  $\Theta$ , la quantité pivotale  $Z_n(\theta) = c_n(T_n - q(\theta))$  converge en loi sous  $P_\theta^n$  vers une variable  $Z(\theta)$  dont la loi  $Q_\theta$  est non dégénérée (non constante). i.e.*

$$\forall \theta \in \Theta, \quad c_n(T_n - q(\theta)) \xrightarrow{P_\theta^n\text{-loi}} Z(\theta),$$

Lorsque  $c_n = \sqrt{n}$  et que la loi limite  $Q_\theta$  est gaussienne, on dit aussi que la suite est **asymptotiquement normale**

*Remarques :*

1. L'interprétation de cette notion est qu'on a alors l'écriture :

$$T_n = q(\theta) + \frac{Z_n(\theta)}{c_n},$$

qui explicite le fait qu'en loi, sous  $P_\theta^n$ ,  $T_n$  converge vers  $q(\theta)$  à la vitesse  $c_n$ .

2. Il est important de remarquer que, dans la définition 21, la quantité pivotale  $Z_n(\theta) = c_n(T_n - q(\theta))$  dépend de  $\theta$ . Ce n'est donc pas un estimateur.
3.  $c_n$  est une "vitesse" au sens où, si  $c_n$  convient  $\lambda c_n$  convient tout aussi bien pour  $\lambda > 0$ .  $\triangle$

Rappelons la définition suivante :

**Définition 22** *Soit  $Z_n$  une suite de variables aléatoires définies sur des espaces (éventuellement) différents, à valeurs dans un espace  $E$ , métrique, de lois  $P_n$ .  $Z$  est une variable aléatoire à valeurs dans  $E$ , de loi  $P$ . On dit que  $Z_n$  converge en loi (sous  $P_n$ ) vers  $Z$  si pour toute fonction  $f$  continue bornée de  $E$  dans  $\mathbf{R}$ ,*

$$Ef(Z_n) = \int f(u)dP_n(u) \longrightarrow Ef(Z) = \int f(u)dP(u)$$

*Remarque :* La mention entre parenthèses sous  $P_n$  est dans ce cadre redondante puisqu'il n'y a pas d'ambiguïté sur la loi de  $Z_n$ . En revanche dans le cadre statistique de la définition 21, cette

précision est fondamentale.  $\triangle$

On rappelle la proposition suivante (cf, par exemple, Billingsley <sup>3</sup>)

**Proposition 11** *Si l'espace d'arrivée  $E$  est  $\mathbf{R}$ , on a équivalence entre les 3 assertions :*

1.  $Z_n$  converge en loi vers  $Z$ .
2.  $E \exp(itZ_n) \longrightarrow E \exp(itZ), \forall t \in \mathbf{R}$ .
3. Pour tout  $x \in \mathbf{R}$  tel que  $P(Z = x) = 0$ ,

$$P_n(Z_n \leq x) \longrightarrow P(Z \leq x).$$

De plus,

1. si  $Z_n$  converge en loi vers  $Z$ , alors pour toute fonction continue  $\phi$ ,  $\phi(Z_n)$  converge en loi vers  $\phi(Z)$ .
2. si  $Z_n \in \mathbf{R}^k$ ,  $Z_n$  converge en loi vers  $Z$  si et seulement si,

$$\forall (t_1, \dots, t_k) \in \mathbf{R}^k, \sum_{i=1}^k t_i Z_n^i \text{ converge en loi vers } \sum_{i=1}^k t_i Z^i$$

## 4.7 Convergence en loi et intervalles de confiance asymptotiques

### 4.7.1 Exemple du modèle de translation échantillonné

**Définition 23**  $\mathcal{E}$  est une expérience de translation si

- $\mathcal{X} = \mathbf{R}$ ,
- $\Theta \subset \mathbf{R}$ ,
- le modèle est dominé par la mesure de Lebesgue sur  $\mathbf{R}$ ,  $\lambda$ ,
- il existe une densité de probabilité  $g$  sur  $\mathbf{R}$  telle que pour tout  $\theta \in \Theta$ ,

$$\frac{dP_\theta}{d\lambda}(x) = g(x - \theta)$$

*Remarque :* Il est facile de montrer que l'observation  $X$  s'écrit alors  $X = \theta + U$  où  $U$  est une variable aléatoire qui admet la loi  $g$  (connue). C'est de cette représentation que vient la dénomination d'expérience de translation.

$\triangle$

Supposons que :

$$\int_{\mathbf{R}} xg(x)dx = 0,$$

$$\int_{\mathbf{R}} x^2g(x)dx = \sigma_g^2 < \infty$$

<sup>3</sup>Billingsley, Patrick, Convergence of probability measures, John Wiley & Sons Inc., New York, 1999

Si l'on veut estimer, dans le modèle échantillonné, la quantité  $q(\theta) = \theta = E_\theta X$ , la méthode des moments propose comme estimateur

$$T_n = \bar{X}_n.$$

Nous avons la proposition suivante :

**Proposition 12** *Dans le modèle de translation échantillonné ci-dessus (la densité  $g$  étant connue), l'estimateur de substitution  $\bar{X}_n$  converge le long de l'expérience échantillonnée  $\mathcal{E}^n$ , à la vitesse  $c_n = \sqrt{n}$ . La variable  $Z(\theta)$  limite admet la loi  $Q_\theta$  normale centrée de variance*

$$V = \sigma_g^2$$

### Démonstration de la Proposition.

C'est une conséquence évidente du théorème de la limite centrale. En effet, sous  $P_\theta^n$  on a,  $Z_n(\theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^n [X_i - \theta] = \frac{1}{\sqrt{n}} \sum_{i=1}^n U_i$ . ■

On en déduit aisément le corollaire

**Corollaire 2** *Dans le modèle de translation échantillonné ci-dessus (la densité  $g$  étant connue) :*

$$[\bar{X}_n - t_{n,\alpha}, \bar{X}_n + t_{n,\alpha}], \quad t_{n,\alpha} = \frac{\sigma_g z_\alpha}{\sqrt{n}}, \quad \int_{z_\alpha}^{\infty} \frac{\exp -\frac{x^2}{2}}{\sqrt{2\pi}} dx = \alpha/2$$

*est un intervalle de confiance asymptotique au niveau  $\alpha$ .*

La démonstration du corollaire est, elle aussi évidente, en remarquant que :

$$\begin{aligned} P_\theta\{\theta \in [\bar{X}_n - t_{n,\alpha}, \bar{X}_n + t_{n,\alpha}]\} &= P_\theta\{\bar{X}_n - \theta \in [-t_{n,\alpha}, +t_{n,\alpha}]\} \\ &= P_\theta\left\{\left|\frac{1}{\sqrt{n}} \sum_{i=1}^n [X_i - \theta]\right| \leq \sigma_g z_\alpha\right\} \\ &= P_\theta\left\{\left|\frac{1}{\sqrt{n}} \sum_{i=1}^n U_i\right| \leq \sigma_g z_\alpha\right\}. \end{aligned}$$

Cette dernière quantité converge vers  $\alpha$ , en utilisant la proposition 11. ■

*Remarque :*

On remarque que la longueur de l'intervalle de confiance construit est exactement  $2t_{n,\alpha}$ , c'est à dire qu'elle est croissante en  $\sigma_g$  et décroissante en fonction du nombre d'observations comme  $1/\sqrt{n}$  (comme dans le cas des intervalles de confiance (exacts) obtenus par l'inégalité d'Hoeffding).  $\triangle$

## 4.8 ICA des méthodes de substitution.

Nous allons maintenant reprendre la démarche de la section précédente obtenue pour le cas simple du modèle de translation, pour étudier le cas général. Nous commençons par démontrer la proposition suivante,

### 4.8.1 Convergence en loi des procédures par substitution

**Proposition 13** Dans un modèle d'échantillonnage, sous les conditions suivantes :

1. Les fonctions  $f_1, \dots, f_r$ , mesurables de  $\mathcal{X}, \mathcal{A}$  dans  $\mathbf{R}, \mathcal{B}(\mathbf{R})$  sont telles que pour tout  $\theta$  dans  $\Theta$  :

$$\mathbf{E}_\theta |f_j(X)|^2 = \int_\Omega |f_j(X)(\omega)|^2 dP_\theta(\omega) < \infty, \forall j \in \{1, \dots, r\}. \quad (4.3)$$

2. la fonction  $g$  est continument différentiable de  $\mathbf{R}^r$  dans  $\mathbf{R}$ , de gradient

$$\nabla g = \left( \frac{\partial g}{\partial x_1}, \dots, \frac{\partial g}{\partial x_r} \right)^*$$

et telle que pour tout  $\theta$  dans  $\Theta$ ,

$$q(\theta) = g(\mathbf{E}_\theta f_1(X), \dots, \mathbf{E}_\theta f_r(X))$$

alors l'estimateur de substitution  $T_n$  converge le long de l'expérience échantillonnée  $\mathcal{E}^n$ , à la vitesse  $c_n = \sqrt{n}$ . La variable  $Z(\theta)$  limite admet la loi

$$Q_\theta = N(0, V(\theta))$$

avec,

$$V(\theta) = \nabla g(\mathbf{E}_\theta f_1(X), \dots, \mathbf{E}_\theta f_r(X))^* \Sigma_\theta \nabla g(\mathbf{E}_\theta f_1(X), \dots, \mathbf{E}_\theta f_r(X))$$

où,

$$(\Sigma_\theta)_{jl} = \text{Cov}_\theta(f_j(X), f_l(X)) = \mathbf{E}_\theta f_j(X) f_l(X) - \mathbf{E}_\theta f_j(X) \mathbf{E}_\theta f_l(X)$$

### Démonstration de la Proposition

Rappelons d'abord le lemme classique suivant :

**Lemme 4 (Slutsky)** Si  $Y_n, A_n, B_n$  sont des suites aléatoires.  $Y_n$  est un vecteur de  $\mathbf{R}^k$ ,  $A_n$  est une matrice de dimension  $l \times k$ ,  $B_n$  est un vecteur de  $\mathbf{R}^l$ . Alors,

1. si  $A_n$  et  $B_n$  convergent en PROBABILITÉ respectivement vers  $A$  et  $B$ , et  $Y_n$  converge en loi vers la variable  $Y$ ,

$$A_n Y_n + B_n$$

converge en loi vers

$$AY + B.$$

2. En particulier si  $A_n$  et  $B_n$  convergent en probabilité vers 0,  $A_n Y_n + B_n$  converge en probabilité vers 0 et (si  $l = k$ )  $Y_n + B_n$  converge en loi vers  $Y$ .

### Démonstration du lemme.

Nous montrons d'abord la partie 2.

1. Pour montrer que  $A_n Y_n + B_n$  tend vers 0 en probabilité, il suffit de considérer le cas où  $B_n = 0$ , en effet la convergence en probabilité passe bien à la somme.

2. Soit  $\eta > 0$  fixé, Pour tout  $\epsilon > 0$ , on a

$$\begin{aligned} P(\|A_n Y_n\| \geq \eta) &\leq P(\|A_n Y_n\| \geq \eta \cap \|A_n\| \leq \epsilon) + P(\|A_n\| \geq \epsilon) \\ &\leq P(\|Y_n\| \geq \frac{\eta}{\epsilon}) + P(\|A_n\| \geq \epsilon) \end{aligned}$$

3. Il suffit maintenant de remarquer en utilisant la proposition 11 :

- (a)  $P(\|A_n\| \geq \epsilon) \rightarrow 0$ , pour tout  $\epsilon > 0$ .
- (b)  $\phi(x) = \|x\|$  étant une fonction continue,  $\|Y_n\|$  converge en loi vers  $\|Y\|$ .
- (c) Nécessairement il existe  $A$  tel que  $P(\|Y\| = A) = 0$ ,  $P(\|Y\| \geq A) \leq \delta$  arbitraire fixé. Prenons alors  $A = \frac{\eta}{\epsilon}$ , nous avons  $P(\|Y_n\| \geq A) \rightarrow P(\|Y\| \geq A)$ , arbitrairement petit.

4. Pour montrer la convergence en loi de  $Y_n + B_n$ , on se ramène en utilisant la dernière assertion de la proposition 11 au cas où les vecteurs sont de dimension 1. Considérons alors,  $x$  tel que  $P(Y = x) = 0$ . Pour tout  $\epsilon$ , tel que  $\|x\| > \epsilon > 0$ , on a

$$\begin{aligned} P(B_n + Y_n \leq x) &\leq P(B_n + Y_n \leq x \cap \|B_n\| \leq \epsilon) + P(\|B_n\| \geq \epsilon) \\ &\leq P(Y_n \leq x + \epsilon) + P(\|B_n\| \geq \epsilon) \end{aligned}$$

$$\begin{aligned} P(B_n + Y_n \leq x) &\geq P(B_n + Y_n \leq x \cap \|B_n\| \leq \epsilon) \\ &\geq P(Y_n \leq x - \epsilon \cap \|B_n\| \leq \epsilon) \\ &\geq P(Y_n \leq x - \epsilon) - P(\|B_n\| > \epsilon) \end{aligned}$$

ce qui montre la convergence en loi en utilisant la proposition 11.

5. Montrons maintenant le passage  $2 \implies 1$  :

Pour cela on considère

- (a)  $(A_n - A)Y_n + B_n - B$ , qui tend en probabilité vers 0 d'après 2.
- (b)  $AY_n + B$  qui tend en loi vers  $AY + B$  en utilisant la proposition 11 et le fait que  $\phi(x) = Ax + B$  est une fonction continue.
- (c) Enfin on écrit  $A_n Y_n + B_n$  comme la somme des 2 termes qu'on vient d'étudier et on utilise la deuxième partie de 2. ■

### Démonstration de la Proposition (suite)

1. Considérons la fonction :

$$\phi(t) = (1-t) \begin{pmatrix} \mathbf{E}_\theta f_1(X) \\ \vdots \\ \mathbf{E}_\theta f_r(X) \end{pmatrix} + t \begin{pmatrix} \hat{f}_1 \\ \vdots \\ \hat{f}_r \end{pmatrix}$$

Il existe  $\tilde{t} \in [0, 1]$ ,

$$\begin{aligned} Z_n(\theta) &= \sqrt{n}(T_n - q(\theta)) \\ &= \sqrt{n}(g \circ \phi(1) - g \circ \phi(0)) \\ &= \sqrt{n}(g \circ \phi)'(\tilde{t}) \\ &= \nabla^* g(\phi(\tilde{t})) \sqrt{n} \begin{pmatrix} \hat{f}_1 - \mathbf{E}_\theta f_1(X) \\ \vdots \\ \hat{f}_r - \mathbf{E}_\theta f_r(X) \end{pmatrix} \\ &= A_n Y_n \end{aligned}$$

où

$$A_n = \nabla^* g \left( \tilde{t}\hat{f}_1 + (1 - \tilde{t})\mathbf{E}_\theta f_1(X), \dots, \tilde{t}\hat{f}_r + (1 - \tilde{t})\mathbf{E}_\theta f_r(X) \right)$$

$$Y_n = \sqrt{n} \begin{pmatrix} \hat{f}_1 - \mathbf{E}_\theta f_1(X) \\ \vdots \\ \hat{f}_r - \mathbf{E}_\theta f_r(X) \end{pmatrix}.$$

2. La condition 1 de la proposition 13, nous permet d'appliquer le théorème de la limite centrale multidimensionnel et d'affirmer que,  $Y_n$  converge en loi sous  $P_\theta^n$ , vers une variable  $Y$  qui suit une loi normale centrée de matrice de covariance  $\Sigma_\theta$ .

3. Comme on a vu précédemment,

$$\forall j \in \{1, \dots, r\}, \hat{f}_j \xrightarrow{P_\theta^n - \text{prob}} \mathbf{E}_\theta f_j(X).$$

Comme  $g$  est continument différentiable, on en déduit que

$$A_n \xrightarrow{P_\theta^n - \text{prob}} \nabla g(\mathbf{E}_\theta f_1(X), \dots, \mathbf{E}_\theta f_r(X))^*.$$

4. La proposition découle alors du lemme de Slutsky.

■

#### 4.8.2 Intervalles de confiance asymptotiques des procédures par substitution.

Si, dans un premier temps nous nous restreignons au cas où  $V(\theta) = V$ , ne dépend pas de  $\theta$ , comme c'est le cas dans le modèle de translation étudié plus haut, nous avons le corollaire suivant.

**Corollaire 3** *Dans le cadre des hypothèses de la Proposition 13, si de plus,  $V(\theta) = V$ , ne dépend pas de  $\theta$ , alors*

$$[T_n - t_{n,\alpha}, T_n + t_{n,\alpha}], \quad t_{n,\alpha} = \frac{\sqrt{V} z_\alpha}{\sqrt{n}}, \quad \int_{z_\alpha}^{\infty} \frac{\exp -\frac{x^2}{2}}{\sqrt{2\pi}} dx = \alpha/2$$

*est un intervalle de confiance asymptotique au niveau  $\alpha$ .*

Malheureusement, dans la plupart des cas,  $V(\theta)$  dépend de  $\theta$  et le corollaire ne s'applique pas.

1. Néanmoins, on peut remarquer que sous les hypothèses de la proposition 13,  $V(\theta)$  est une quantité de la forme  $q_1(\theta)$ , estimable par méthode de substitution, à l'aide des fonctions  $f_i$ ,  $i \in \{1, \dots, r\}$  et  $f_i f_j$ ,  $i, j \in \{1, \dots, r\}$ .
2. Cette remarque nous permet de construire un estimateur par substitution de  $V(\theta)$  :

$$V_n$$

qui est convergent de par les hypothèses qu'on a fait dans le cadre de l'utilisation de la proposition 9.

Ceci permet d'établir la proposition suivante :

**Proposition 14** *Dans le cadre des hypothèses de la Proposition 13, alors*

$$[T_n - t_{n,\alpha}, T_n + t_{n,\alpha}], \quad t_{n,\alpha} = \frac{\sqrt{V_n} z_\alpha}{\sqrt{n}}, \quad \int_{z_\alpha}^{\infty} \frac{\exp -\frac{x^2}{2}}{\sqrt{2\pi}} dx = \alpha/2$$

*est un intervalle de confiance asymptotique au niveau  $\alpha$ .*

La démonstration est élémentaire en utilisant le résultat de la proposition 13, la convergence de  $V_n$  vers  $V(\theta)$  en  $P_\theta^n$  probabilité et le lemme de Slutski, pour affirmer que

$$\sqrt{nV_n^{-1}}(T_n - q(\theta))$$

converge en loi sous  $P_\theta^n$  vers une variable gaussienne centrée réduite.

*Exercice :* En reprenant les notations du paragraphe 4.7.1 on désigne par  $P_\theta$ , la loi de la variable aléatoire  $\mu + U$  où  $U$  suit la loi de densité  $g$ . On a toujours  $\int xg(x)dx = 0$  et  $\int x^2g(x)dx = \sigma^2$ . Mais cette fois,  $\sigma^2$  est inconnu. Maintenant  $\theta = (\mu, \sigma^2)$  et on supposera de plus que  $\int x^4g(x)dx < \pm\infty$ . Construire des ICA( $\alpha$ ) pour resp. les quantités  $q_1(\theta) = \mu$  et  $q_2(\theta) = \sigma^2$ .



# Chapitre 5

## METHODE DE CONTRASTES.

### 5.1 Méthodes de contraste, définitions

Le cadre principal sera souvent, comme le chapitre précédent, une expérience échantillonnée  $\mathcal{E}^n$  d'une expérience primitive

$$\mathcal{E} = (X, \mathcal{X}, \mathcal{A}, P_\theta, \theta \in \Theta),$$

mais nous considérerons aussi le modèle linéaire. On va principalement s'occuper d'estimer la quantité  $q(\theta) = \theta$ .

**Définition 24** On appelle fonction de contraste sur  $\Theta$ , toute fonction

$$\gamma : \Theta \times \Theta \longrightarrow \mathbf{R}$$

telle que pour tout  $\theta \in \Theta$ , la fonction,

$$\alpha \longrightarrow \gamma(\theta, \alpha)$$

admet un minimum unique en

$$\alpha = \theta$$

*Exemples :*

1. si  $\Theta$  est un sous ensemble de  $\mathbf{R}^d$ , si  $A$  est une matrice symétrique définie positive, alors

$$\gamma(\theta, \alpha) = \frac{1}{2}(\theta - \alpha)^* A(\theta - \alpha)$$

est une fonction de contraste.

2. Définissons la quantité suivante pour  $P$  et  $Q$ , 2 probabilités définies sur le même espace, et qui porte le nom **d'Information de Küllback** :

$$K(P, Q) := +\infty \text{ si } P \text{ n'est pas dominée par } Q$$

$$\int \frac{dP}{dQ} \log \frac{dP}{dQ} dQ, \text{ si } P \text{ est dominée par } Q$$

On montre aisément, en appliquant l'inégalité de Jensen à la fonction  $\phi(x) = x \log x$  (strictement convexe) que  $K(P, Q) \geq 0$  et  $K(P, Q) = 0$  si et seulement si  $p = \frac{dP}{dQ}$  est presque sûrement constant, c'est à dire si  $P = Q$ . On en déduit que sur une expérience  $\mathcal{E}$ , la quantité

$$K(\theta, \alpha) = K(P_\theta, P_\alpha)$$

est une fonction de contraste.

△

**Définition 25**  $\mathcal{E}_n = (X^n, \mathcal{X}_n, \mathcal{A}_n, P_\theta^n, \theta \in \Theta)$ , étant une suite générale d'expériences, on appelle processus de contraste associé à la fonction de contraste  $\gamma$  une suite de fonctions aléatoires adaptée à  $\mathcal{E}_n$

$$\alpha \in \Theta \mapsto U_n(\alpha, X^n) \in \mathbf{R}$$

telle pour tout  $\theta, \alpha \in \Theta$ ,

$$U_n(\alpha, X^n) - U_n(\theta, X^n) \xrightarrow{P_\theta^n - \text{prob}} \gamma(\theta, \alpha)$$

L'idée maintenant consiste à remarquer que la définition précédente nous suggère que sous  $P_\theta^n$ ,  $U_n(\alpha, X^n) - U_n(\theta, X^n)$  est proche d'une fonction de contraste. Considérée comme fonction de  $\alpha$ , cette fonction nous permet donc de repérer  $\theta$  en cherchant l'endroit où elle atteint son minimum. Bien entendu, ceci ne mènerait à rien (puisque la fonction dépend de  $\theta$ ) si l'on ne faisait la remarque simple suivante :

Minimiser  $U_n(\alpha, X^n) - U_n(\theta, X^n)$ , équivaut à minimiser  $U_n(\alpha, X^n)$  fonction qui elle ne dépend pas de  $\theta$ .

On en déduit la définition suivante :

**Définition 26** Dans le contexte ci-dessus, on appelle estimateur de contraste associé l'estimateur  $T_n \in \Theta$ , quand il existe et est unique qui vérifie :

$$\forall \theta \in \Theta, \quad U_n(T_n, X^n) \leq U_n(\theta, X^n)$$

Cette définition nous permet de donner comme cas particuliers 2 catégories importantes de méthodes d'estimation : L'estimateur des moindres carrés dans le modèle linéaire et l'estimateur du maximum de vraisemblance, dans une expérience dominée.

*Remarque :*

Si au lieu d'estimer  $\theta$  on estime  $q(\theta)$  on définit alors l'estimateur de contraste associé par simple 'plug-in' :

$$\hat{q}_n = q(T_n)$$

△

## 5.2 Estimateur du maximum de vraisemblance

Considérons une expérience

$$\mathcal{E} = (X, \mathcal{X}, \mathcal{A}, P_\theta, \theta \in \Theta),$$

dominée par une mesure  $\mu$ . On va supposer en outre que les mesures  $P_\theta$  sont toutes équivalentes à  $\mu$ , soit encore que

$$p_\theta(x) := \frac{dP_\theta}{d\mu}(x) > 0, \quad \forall x$$

On considère  $\mathcal{E}^n$  échantillonnée de  $\mathcal{E}$ . Considérons le processus

$$U_n(\theta, X^n) = -\frac{1}{n} \sum_{i=1}^n \log \frac{dP_\theta}{d\mu}(X_i)$$

On vérifie aisément que si pour tout  $\theta, \alpha \in \Theta$ ,

$$\mathbf{E}_\theta \left| \log \frac{dP_\alpha}{d\mu}(X) \right| < \infty$$

la loi des grands nombres implique :

$$\begin{aligned} U_n(\alpha, X^n) - U_n(\theta, X^n) &\xrightarrow{P_\theta^n\text{-prob}} -\mathbf{E}_\theta \log \frac{dP_\alpha}{d\mu}(X) + \mathbf{E}_\theta \log \frac{dP_\theta}{d\mu}(X) \\ &= \mathbf{E}_\theta \log \frac{dP_\theta}{dP_\alpha}(X) \\ &= K(P_\theta, P_\alpha) \end{aligned}$$

qui est notre deuxième exemple de fonction de contraste. Ceci nous permet, en élargissant un peu le contexte précédent de définir la notion importante suivante.

**Définition 27** Dans une expérience dominée par une mesure  $\mu$ , échantillonnée, on appelle **estimateur du maximum de vraisemblance** l'estimateur  $T_n \in \Theta$ , quand il existe qui vérifie :

$$\begin{aligned} \text{Pour } L_n(\theta, X^n) &= \frac{1}{n} \sum_{i=1}^n \log \frac{dP_\theta}{d\mu}(X_i), \\ \forall \theta \in \Theta, L_n(T_n, X^n) &\geq L_n(\theta, X^n) \end{aligned}$$

### 5.3 Estimateur des moindres carrés

Considérons maintenant un modèle linéaire, non nécessairement gaussien : On observe :

$$Y^n = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, M(n) = \begin{pmatrix} M_{11}(n) & \dots & M_{1p}(n) \\ & \ddots & \\ M_{n1}(n) & \dots & M_{np}(n) \end{pmatrix},$$

sous  $P_\theta^n$ ,  $Y^n = M(n)\theta + \varepsilon$ ,  $\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$ .

où les  $\varepsilon_i$  sont des variables indépendantes, de même loi  $g$ , connue, centrée et qui possède un moment d'ordre 2. (Cette loi est une gaussienne centrée, dans le modèle linéaire gaussien.) Considérons le processus pour  $\theta \in \mathbf{R}^p$

$$U_n(\theta, Y^n) = \frac{1}{2n} \sum_{i=1}^n (Y_i - \sum_{j=1}^p M_{ij}(n)\theta_j)^2$$

Sous  $P_\theta^n$ ,

$$\begin{aligned}
U_n(\alpha, X^n) - U_n(\theta, X^n) &= \frac{1}{2n} \sum_{i=1}^n [Y_i - \sum_{j=1}^p M_{ij}(n)\theta_j + \sum_{j=1}^p M_{ij}(n)(\theta_j - \alpha_j)]^2 \\
&- \frac{1}{2n} \sum_{i=1}^n (Y_i - \sum_{j=1}^p M_{ij}(n)\theta_j)^2 \\
&= \frac{1}{2n} \sum_{i=1}^n [\epsilon_i + \sum_{j=1}^p M_{ij}(n)(\theta_j - \alpha_j)]^2 - \frac{1}{2n} \sum_{i=1}^n (\epsilon_i)^2 \\
&= \frac{1}{n} \sum_{i=1}^n \epsilon_i [\sum_{j=1}^p M_{ij}(n)(\theta_j - \alpha_j)] + \frac{1}{2n} \sum_{i=1}^n [\sum_{j=1}^p M_{ij}(n)(\theta_j - \alpha_j)]^2
\end{aligned}$$

Posons  $R(n) = \frac{1}{n} M(n)^* M(n)$ , et faisons l'hypothèse suivante (remarquons que pour tout  $n$ , la dimension de  $R(n)$  est toujours  $p \times p$ ).

$\mathcal{H}$  : Il existe  $A$  matrice symétrique définie positive telle  $R(n) \xrightarrow{n \rightarrow \infty} A$

Alors,

$$U_n(\alpha, Y^n) - U_n(\theta, Y^n) \xrightarrow{P_\theta^n - prob} \frac{1}{2} (\theta - \alpha)^* A (\theta - \alpha)$$

qui est notre premier exemple de fonction de contraste.

En effet, on vérifie aisément que :

1.

$$\frac{1}{2n} \sum_{i=1}^n [\sum_{j=1}^p M_{ij}(n)(\theta_j - \alpha_j)]^2 = \frac{1}{2} (\theta - \alpha)^* R(n) (\theta - \alpha) \xrightarrow{P_\theta^n - prob} \frac{1}{2} (\theta - \alpha)^* A (\theta - \alpha).$$

2. Par ailleurs, si  $Z_n = \frac{1}{n} \sum_{i=1}^n \epsilon_i [\sum_{j=1}^p M_{ij}(n)(\theta_j - \alpha_j)]$ , on a :

$$\begin{aligned}
\mathbf{E}_\theta Z_n = 0, \text{ Var}_\theta Z_n &= \frac{1}{n^2} \sum_{i=1}^n \text{Var}_{\epsilon_i} [\sum_{j=1}^p M_{ij}(n)(\theta_j - \alpha_j)]^2 \\
&= \frac{1}{n} (\theta - \alpha)^* R(n) (\theta - \alpha) \text{Var}_{\epsilon_i} \\
&\rightarrow 0
\end{aligned}$$

Et donc,  $Z_n \xrightarrow{P_\theta^n - prob} 0$ .

Ceci nous permet, en élargissant un peu le contexte, de définir la notion importante suivante.

**Définition 28** Dans le modèle linéaire général suivant

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, M = \begin{pmatrix} M_{11} & \dots & M_{1p} \\ \vdots & & \vdots \\ M_{n1} & \dots & M_{np} \end{pmatrix}, Y = M\beta + \epsilon, \epsilon = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

où les  $\epsilon_i$  sont des variables indépendantes, de même loi  $g$ , connue, centrée et qui possède un moment d'ordre 2, on appelle **estimateur des moindres carrés** l'estimateur  $\hat{\beta} \in \mathbf{R}^p$ , qui vérifie :

$$\text{Pour } \gamma_n(\beta, Y) = \sum_{i=1}^n (Y_i - \sum_{j=1}^p M_{ij}\beta_j)^2$$

$$\forall \beta \in \mathbf{R}^p, \quad \gamma_n(T_n, Y) \leq \gamma_n(\beta, Y)$$

*Exercice :* Montrer que les 2 notions précédentes d'estimateur du maximum de vraisemblance et des moindres carrés coïncident dans le cadre du modèle linéaire gaussien.  $\triangle$

## 5.4 Calculs d'estimateurs du maximum de vraisemblance

### 5.4.1 Problèmes liés à la définition de l'estimateur.

Montrons d'abord que la définition précédente de l'estimateur du maximum de vraisemblance peut poser un certain nombre de problèmes :

1. L'estimateur peut ne pas exister :

Prenons l'exemple simple d'un modèle gaussien.  $\theta = (\mu, \sigma^2) \in \Theta = \mathbf{R} \times \mathbf{R}_+^*$ ,  $P_\theta = N(\mu, \sigma^2)$ . Il est facile de montrer que le modèle est dominé par la mesure de Lebesgue, de densité :

$$p_\theta(x) = \frac{1}{(2\pi\sigma)^{1/2}} \exp \frac{-x^2}{2\sigma^2}$$

On voit facilement, que pour  $n = 1$ , maximiser  $L_1$  revient à minimiser la quantité :

$$\log \sigma + \frac{(X_1 - \mu)^2}{2\sigma^2}$$

c'est à dire à prendre  $\hat{\mu} = X_1$ ,  $\hat{\sigma} = 0$ , qui n'appartient pas à  $\Theta$  !

2. Il peut aussi ne pas être unique :

Prenons maintenant l'exemple d'un modèle uniforme.  $\Theta = \mathbf{R}_+^*$ ,  $P_\theta$  est la loi uniforme sur  $[\theta, \theta + 1]$ . Il est facile de montrer que le modèle est dominé par la mesure de Lebesgue, de densité :

$$p_\theta(x) = I\{[\theta, \theta + 1]\}(x).$$

On voit facilement, que maximiser  $L_n$  revient à maximiser la quantité :

$$I\{[\inf(X_i), \sup(X_i) - 1]\}(\theta).$$

Or cette quantité est clairement égale à 1 (et donc maximum) pour toute valeur comprise entre  $\inf(X_i)$  et  $\sup(X_i) - 1$ .

3. On peut aussi se poser la question de la dépendance de l'estimateur par rapport à la mesure dominante  $\mu$ . En fait, nous avons la proposition :

**Proposition 15** *La définition de l'estimateur du maximum de vraisemblance ne dépend pas de la mesure dominante choisie.*

**Preuve de la Proposition** Soit  $\mu_1$  et  $\mu_2$  2 mesures dominantes, posons  $\mu^* = \mu_1 + \mu_2$ , comme

$$\frac{dP_\theta}{d\mu^*}(x) = \frac{dP_\theta}{d\mu_1}(x) \frac{d\mu_1}{d\mu^*}(x) = \frac{dP_\theta}{d\mu_2}(x) \frac{d\mu_2}{d\mu^*}(x).$$

On voit sur cette expression que la maximisation en  $\theta$  de chacune de ces quantités va donner le même résultat.

4. Donnons toutefois un exemple de calcul de l'estimateur du maximum de vraisemblance où il n'y a pas de difficultés. Reprenons l'exemple simple du modèle gaussien.  $\theta = (\mu, \sigma^2) \in \Theta = \mathbf{R} \times \mathbf{R}_+^*$ ,  $P_\theta = N(\mu, \sigma^2)$ . Prenons maintenant  $n > 1$ , maximiser  $L_n$  revient, en posant  $\tau = \sigma^2$  à maximiser la quantité :

$$l_n(\theta) = -\frac{n}{2} \log \tau - \frac{\sum_{i=1}^n (X_i - \mu)^2}{2\tau}.$$

Si on calcule le gradient en  $(\mu, \tau)$ , de cette quantité, on trouve

$$\nabla l_n(\theta) = \left( \begin{array}{c} \sum_{i=1}^n \frac{(X_i - \mu)}{\tau} \\ -\frac{n}{2\tau} + \sum_{i=1}^n \frac{(X_i - \mu)^2}{2\tau^2} \end{array} \right)$$

Ce gradient s'annule en un seul point

$$\left( \begin{array}{c} \hat{\mu} \\ \hat{\tau} \end{array} \right) = \left( \begin{array}{c} \bar{X}_n \\ \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2 \end{array} \right)$$

Une étude simple de la fonction  $l_n(\theta)$  montre que ce point est un maximum non seulement local mais aussi global.

#### 5.4.2 Calcul itératif et maximum de vraisemblance

L'une des difficultés soulevées par cette méthode d'estimation est souvent la difficulté de sa mise en oeuvre. Donnons un exemple simple.

**Observation complète** Supposons que nous observions les temps de panne  $T_1, \dots, T_n$  de  $n$  objets indépendants. On suppose qu'ils ont tous même loi, exponentielle de paramètre  $\theta > 0$ . i.e.

$$P_\theta(T_i \leq t) = \int_0^t \theta \exp \theta u \, du = 1 - \exp -\theta t.$$

Dans ce cadre le calcul de l'estimateur du maximum de vraisemblance ne pose pas non plus de problème et donne :

$$\hat{\theta} = \frac{n}{\sum_{i=1}^n T_i}.$$

**Observation partielle** Supposons, qu'au lieu d'observer les  $T_i$ , nous ne disposions que d'observations partielles (qui correspondent en pratique au fait qu'au lieu de surveiller en permanence les objets pour noter l'heure exacte de leur panne, on fait des rondes, par exemple toutes les heures), les  $Y_i$  où

$$\begin{aligned} Y_i &= l \quad \text{si } l-1 < T_i \leq l, \quad l = 1, \dots, k-1 \\ &= k+1 \quad \text{si } k < T_i \end{aligned}$$

On montre facilement que dans ce nouveau modèle, l'estimateur du maximum de vraisemblance rend maximum la quantité :

$$\prod_{j=1}^{k+1} p_j(\theta)^{N_j} \quad \text{avec :}$$

$$\begin{aligned}
p_l(\theta) &= P_\theta(Y_i = l) = \exp(-(l-1)\theta) - \exp(-l\theta), \quad l = 1, \dots, k_1 \\
p_{k+1}(\theta) &= P_\theta(Y_i = k+1) = \exp(-k\theta) \\
N_j &= \sum_{i=1}^n I\{Y_i = j\}
\end{aligned}$$

On peut montrer que cette expression admet un maximum, mais que ce maximum n'a pas d'expression explicite. Nous allons donc le rechercher de façon approchée. Il existe de nombreuses méthodes pour faire cela, y compris des méthodes stochastiques. Nous allons ici expliciter une méthode très simple, la méthode de Newton Raphson.

**Calcul approché de l'estimateur du maximum de vraisemblance** Les bases de ce calcul sont les suivantes :

1. On se propose de rechercher l'argument du maximum sur un ouvert  $I$  de  $\mathbf{R}^d$  de la fonction

$$\theta \in I \longrightarrow L_n(\theta, X^n).$$

2. On suppose que cette fonction est au moins 2 fois continument différentiable et qu'elle est strictement concave de sorte qu'elle n'admet qu'un seul extremum local qui est son maximum.
3. On va donc chercher une solution du système d'équations :

$$(E) : G_n(\theta, X^n) = \nabla L_n(\theta, X^n) = 0.$$

Où  $\nabla L_n(\theta, X^n)$  désigne le gradient en  $\theta$  de  $L_n(\theta, X^n)$ .

4. Pour cela on propose l'algorithme itératif suivant :

(a) Soit  $\hat{\theta}_k$  la valeur à la  $k$ ième étape. Si  $G_n(\hat{\theta}_k, X^n) = 0$ , on s'arrête. Sinon,

(b) On pose

$$(E') : \hat{\theta}_{k+1} = \hat{\theta}_k + H_n(\hat{\theta}_k, X^n)^{-1} G_n(\hat{\theta}_k, X^n)$$

où  $H_n(\theta, X^n)$  désigne la matrice des dérivées seconde en  $\theta$  de  $L_n(\theta, X^n)$

(c) On itère.

(d) On se fixe un pas d'arrêt  $\eta$  et on arrête l'algorithme dès que  $\|\hat{\theta}_{k+1} - \hat{\theta}_k\| \leq \eta$ .

*Remarques :*

1. L'équation  $(E')$  est obtenue à partir de l'équation  $(E)$  par un simple développement de Taylor au premier ordre (autour du point  $\hat{\theta}_k$ ) dont on néglige le reste.
2. Il est clair que sous les hypothèses que l'on a faites, la quantité  $H_n(\hat{\theta}_k, X^n)^{-1}$  est toujours définie.
3. On montre que sous les hypothèses que l'on a faites (et même dans un cadre plus large) l'algorithme présenté converge vers le maximum cherché, pourvu que  $\hat{\theta}_0$ , l'initialisateur soit bien choisi.
4. Ce choix influe aussi sur la rapidité de l'algorithme, et sur les propriétés de convergence en loi de l'estimateur final obtenu. On choisit généralement pour  $\hat{\theta}_0$ , un estimateur préliminaire, convergent et facile à calculer, par exemple obtenu par méthode de substitution.

$\triangle$

## 5.5 Calcul et lois des estimateurs des moindres carrés

Remarquons d'abord que la fonction

$$\gamma_n(\beta, Y) = \sum_{i=1}^n (Y_i - (M\beta)_i)^2$$

mesure la distance dans  $\mathbf{R}^n$  entre le vecteur  $Y$  et sa prédiction par  $M\beta$ . Il est relativement naturel de choisir comme estimateur de  $\beta$ , un point  $\hat{\beta}$  rendant cette quantité minimum.

$$\hat{\beta} = \text{Argmin}\{\gamma_n(\beta, Y); \beta \in \mathbf{R}^p\}$$

### 5.5.1 Interprétation géométrique

Si  $\beta$  parcourt  $\mathbf{R}^p$ ,  $M\beta$  parcourt l'espace vectoriel  $V$  engendré, dans  $\mathbf{R}^n$ , par les colonnes de la matrice  $M$  :

$$V = M(\mathbf{R}^p) \subset \mathbf{R}^n$$

Comme  $\gamma_n(\beta, Y) = \|Y - M\beta\|^2$ , nécessairement  $M\hat{\beta}$ , existe, est unique puisque c'est la projection sur  $V$  de  $Y$ ,  $M\hat{\beta} = \text{Proj}_V(Y)$ . On en déduit que  $\hat{\beta}$  existe aussi toujours, mais n'est unique que si  $M$  est injectif.

**Proposition 16** *Si  $p \leq n$ , la matrice  $M$ , de dimension  $n \times p$  est injective si et seulement si  $M^*M$  est inversible.*

#### Démonstration de la Proposition.

Il suffit de démontrer que  $\ker(M) = \ker(M^*M)$ . Il est clair que  $\ker(M) \subset \ker(M^*M)$ . Maintenant, soit  $u \in \ker(M^*M)$ , on a  $M^*Mu = 0$ , d'où  $u^*M^*Mu = 0$ , i.e.  $\|Mu\|^2 = 0 \implies Mu = 0 \implies u \in \ker M$ .

#### Résolution algébrique

$$\begin{aligned} M\hat{\beta} = \text{Proj}_V(Y) &\iff \langle Y - M\hat{\beta}, Mb \rangle = 0, \quad \forall b \in \mathbf{R}^p \\ &\iff b^*M^*Y = b^*M^*M\hat{\beta}, \quad \forall b \in \mathbf{R}^p \\ &\iff M^*Y = M^*M\hat{\beta} \end{aligned}$$

D'où, en utilisant la proposition si  $M$  est injective,

$$\hat{\beta} = (M^*M)^{-1}M^*Y$$

*Remarque :* Si  $M^*M$  n'est pas inversible, on n'a pas unicité de  $\hat{\beta}$ , mais existence. Donnons une solution, utilisant la pseudoinverse :  $M^*M$  étant une matrice symétrique, positive, elle s'écrit  $R^*DR$  avec  $R$  matrice orthogonale et  $D$  est une matrice diagonale, dont les coefficients diagonaux sont notés  $r_i^2$ . On suppose  $r_i^2 > 0, \forall i = 1, \dots, k$ ,  $r_i^2 = 0, \forall i \geq k + 1$ . Appelons

pseudoinverse de  $M^*M$  la matrice

$$(M^*M)^{(-1*)} = R^* \begin{pmatrix} \frac{1}{r_1^2} & \dots & 0 & 0 & \dots & 0 \\ & & \vdots & & & \\ 0 & \dots & \frac{1}{r_k^2} & 0 & \dots & 0 \\ & & \vdots & & & \\ 0 & \dots & 0 & 0 & \dots & 0 \end{pmatrix} R$$

Notons que si  $M^*M$  est inversible, alors pseudoinverse et inverse coïncident. On vérifie facilement que

$$\hat{\beta} = (M^*M)^{(-1*)}M^*Y$$

est une solution de notre problème, et que l'opérateur de projection sur  $V$  est donné par :

$$M\hat{\beta} = M(M^*M)^{(-1*)}M^*Y = \text{Proj}_V(Y)$$

△

Rappelons que si  $V^\perp$  est le supplémentaire orthogonal de  $V$ ,

$$\text{Proj}_{V^\perp}(Y) = Y - \text{Proj}_V(Y) = [I_n - \text{Proj}_V](Y) = [I_n - M(M^*M)^{-1}M^*]Y$$

**Définition 29** On appelle vecteur des résidus, le vecteur

$$\hat{\varepsilon} = [I_n - M(M^*M)^{-1}M^*]Y.$$

Il représente l'erreur de prédiction. Le carré de sa norme s'appelle l'erreur quadratique.

*Exemples :*

1. Dans le cas élémentaire suivant :

$$Y_i = \mu + \varepsilon_i$$

l'estimateur des moindres carrés se calcule facilement et vaut  $\bar{Y}_n = \frac{\sum_{i=1}^n Y_i}{n}$ .

2. Dans le cas d'une régression linéaire, nous avons vu que  $\beta = (a, b)^*$  et

$$M = \begin{pmatrix} M_1 & 1 \\ \vdots & \vdots \\ M_n & 1 \end{pmatrix}$$

De sorte que

$$M^*M = \begin{pmatrix} \sum_{i=1}^n M_i^2 & \sum_{i=1}^n M_i \\ \sum_{i=1}^n M_i & n \end{pmatrix}$$

Dans ce cas, un changement de paramètres peut rendre les choses plus aisées : En effet, si on introduit  $\bar{M}_n = \frac{\sum_{i=1}^n M_i}{n}$ , le modèle s'écrit :

$$Y_i = az_i + b' + \varepsilon_i, \quad z_i = M_i - \bar{M}_n, \quad b' = b + a\bar{M}_n \quad (5.1)$$

et clairement minimiser en  $a, b'$ ,

$$\sum_{i=1}^n (Y_i - az_i + b')^2$$

équivalent à minimiser en  $a, b$ ,

$$\sum_{i=1}^n (Y_i - aM_i + b)^2$$

avec la relation suivante  $\hat{b}' = \hat{b} + \hat{a}\bar{M}_n$ . L'équation (5.1) introduit un nouveau modèle linéaire dont la matrice endogène  $M'$  s'écrit :

$$M'^*M' = \begin{pmatrix} \sum_{i=1}^n z_i^2 & 0 \\ 0 & n \end{pmatrix}$$

Cette matrice est inversible si et seulement si  $\sum_{i=1}^n z_i^2 \neq 0$ , c'est à dire si les  $M_i$  ne sont pas tous égaux. Dans ce cas, on obtient facilement :

$$\hat{a} = \frac{\sum_{i=1}^n (M_i - \bar{M}_n)Y_i}{\sum_{i=1}^n (M_i - \bar{M}_n)^2}, \quad \hat{b} = \bar{Y}_n + \hat{a}\bar{M}_n$$

3. Considérons maintenant la régression périodique suivante :

$$Y_i = a_0 + a_1 \cos(2\pi \frac{i}{n}) + a_2 \sin(2\pi \frac{i}{n}) + \varepsilon_i, \quad i = 1, \dots, n$$

On vérifie, en utilisant les relations sur les racines de l'unité que  $M^*M$  se met sous la forme suivante :

$$\begin{pmatrix} n & \sum_{i=1}^n \cos(2\pi \frac{i}{n}) & \sum_{i=1}^n \sin(2\pi \frac{i}{n}) \\ \sum_{i=1}^n \cos(2\pi \frac{i}{n}) & \sum_{i=1}^n \cos(2\pi \frac{i}{n})^2 & \sum_{i=1}^n \cos(2\pi \frac{i}{n}) \sin(2\pi \frac{i}{n}) \\ \sum_{i=1}^n \sin(2\pi \frac{i}{n}) & \sum_{i=1}^n \cos(2\pi \frac{i}{n}) \sin(2\pi \frac{i}{n}) & \sum_{i=1}^n \sin(2\pi \frac{i}{n})^2 \end{pmatrix} = \begin{pmatrix} n & 0 & 0 \\ 0 & \frac{n}{2} & 0 \\ 0 & 0 & \frac{n}{2} \end{pmatrix}$$

On en déduit que

$$\hat{a}_0 = \bar{Y}_n, \quad \hat{a}_1 = \frac{1}{2n} \sum_{i=1}^n \cos(2\pi \frac{i}{n}) Y_i, \quad \hat{a}_2 = \frac{1}{2n} \sum_{i=1}^n \sin(2\pi \frac{i}{n}) Y_i$$

△

### 5.5.2 Lois des estimateurs dans le cas gaussien. Estimation de $\sigma^2$ .

Plaçons nous maintenant dans le cas où le vecteur  $\epsilon$  est gaussien  $N(0, \sigma^2 I_n)$ . Nous allons maintenant montrer la proposition suivante :

**Proposition 17** *Sous la condition,  $M^*M$  inversible, le vecteur de dimension  $p + n$  :*

$$\begin{pmatrix} \hat{\beta} \\ \hat{\epsilon} \end{pmatrix}$$

est un vecteur gaussien de moyenne et variance :

$$\begin{pmatrix} \beta \\ 0 \end{pmatrix}, \quad \sigma^2 \begin{pmatrix} (M^*M)^{-1} & 0 \\ 0 & I_n - M(M^*M)^{-1}M^* \end{pmatrix}$$

**Preuve de la Proposition****Espérances et variances de  $\hat{\beta}$  :**

Notons d'abord que dans ce paragraphe, l'hypothèse de gaussiannité sur les  $\varepsilon_i$  est inutile. Les résultats sont encore vrais si l'on suppose que  $\mathbf{E}\varepsilon = 0$ ,  $\text{Var}\varepsilon = \sigma^2 I_n$ .

Comme  $\hat{\beta} = (M^*M)^{-1}M^*Y$ , on a  $\mathbf{E}\hat{\beta} = \mathbf{E}(M^*M)^{-1}M^*(M\beta + \varepsilon) = \beta$ .

D'autre part,

$$\begin{aligned}\text{Var}(\hat{\beta}) &= (M^*M)^{-1}M^*[\text{Var}(Y)]M(M^*M)^{-1} \\ &= (M^*M)^{-1}M^*[\text{Var}(\varepsilon)]M(M^*M)^{-1} \\ &= \sigma^2(M^*M)^{-1}M^*M(M^*M)^{-1} = \sigma^2(M^*M)^{-1}.\end{aligned}$$

**Loi du vecteur**

1. Le vecteur

$$\begin{pmatrix} \hat{\beta} \\ \hat{\varepsilon} \end{pmatrix}$$

est fonction linéaire du vecteur  $Y$ , c'est donc un vecteur gaussien.

2.  $\hat{\beta} \sim N(\beta, \sigma^2(M^*M)^{-1})$  :

Nous avons calculé la moyenne et la variance de  $\hat{\beta}$  au paragraphe précédent. Sa loi est donc déterminée.

3.  $\hat{\beta}$  et  $\hat{\varepsilon}$  sont indépendants :

Il est immédiat que  $\mathbf{E}\hat{\varepsilon} = 0$ .

Nous avons vu que :  $M\hat{\beta} = \text{Proj}_V(Y) = M\beta + e$  avec  $e = \text{Proj}_V(\varepsilon)$ .

De plus,  $\hat{\varepsilon} = [I_n - \text{Proj}_V](Y) = \text{Proj}_{V^\perp}(Y) = \text{Proj}_{V^\perp}(\varepsilon) = \varepsilon - e$ .

Soit maintenant  $P_1 = \text{Proj}_V = M(M^*M)^{-1}M^*$  et  $P_2 = \text{Proj}_{V^\perp} = I_n - M(M^*M)^{-1}M^*$ .

On a donc  $M\hat{\beta} = M\beta + P_1\varepsilon$ ,  $\hat{\varepsilon} = P_2\varepsilon$ .

Par ailleurs,  $P_1 + P_2 = I_n$ ,  $\text{rg}(P_1) = \dim V = \text{rg}M = p$ ,  $\text{rg}(P_2) = n - p$ .

On peut donc appliquer le th'eorème de Cochran et en déduire que  $e$  et  $\hat{\varepsilon}$  sont indépendants.

Par conséquent,  $M\hat{\beta}$  et  $\hat{\varepsilon}$  sont indépendants.

Il en est de même pour  $M^*M\hat{\beta}$  et  $\hat{\varepsilon}$ , et donc pour  $\hat{\beta}$  et  $\hat{\varepsilon}$ .

4. Il nous reste à calculer la matrice de covariance du vecteur  $\hat{\varepsilon}$ .

Mais, comme  $\hat{\varepsilon} = P_2\varepsilon$ , elle est égale à  $\sigma^2 P_2$ .

Ceci achève la preuve de la proposition. ■

**Estimation de  $\sigma^2$ .**

En appliquant le résultat de la Proposition 8, nous avons :

$\|\hat{\varepsilon}\|^2$  suit une loi  $\sigma^2\chi^2(n-p)$ . En conséquence,  $\hat{\sigma}^2 = \frac{\|\hat{\varepsilon}\|^2}{n-p}$  est d'espérance  $\sigma^2$ . C'est donc un estimateur assez naturel de  $\sigma^2$ .

**En résumé :**

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

De plus ces 2 estimateurs sont indépendants.

### 5.5.3 Intervalles de confiance pour $a^*\beta$ et $\sigma^2$

Soit  $a^*$  un vecteur connu de  $(\mathbf{R}^p)^*$ . On se propose d'estimer  $a^*\beta$ .

*Exemples :*

1. Si  $a^* = (1, 0, \dots, 0)$ , on s'intéresse à estimer  $\beta_1$ .
2. Dans l'exemple d'une comparaison de 2 populations,  $p = 2$ , prendre  $a^* = (1, -1)$  consiste à estimer la différence des moyennes.

△

On va prendre naturellement  $a^*\hat{\beta}$  comme estimateur de  $a^*\beta$ . Nous nous proposons de construire un intervalle de confiance associé à cette estimation.

#### Estimation de $a^*\beta$ , $\sigma^2$ étant connu

On vérifie que  $a^*(\hat{\beta} - \beta) \sim N(0, \sigma^2 a^*(M^*M)^{-1}a)$ , de sorte que si  $\Phi(z_{\alpha/2}) = \alpha/2$ ,

$$[a^*\hat{\beta} - z_{\alpha/2}\sqrt{a^*(M^*M)^{-1}a}\sigma, a^*\hat{\beta} + z_{\alpha/2}\sqrt{a^*(M^*M)^{-1}a}\sigma]$$

est un intervalle de confiance pour la quantité  $a^*\beta$ , au niveau d'erreur  $\alpha$ .

#### Estimation de $a^*\beta$ , $\sigma^2$ étant inconnu

On a

1.  $a^*(\hat{\beta} - \beta) \sim N(0, \sigma^2 a^*(M^*M)^{-1}a)$ ,
2.  $\hat{\sigma}^2 \sim \frac{\sigma^2}{n-p}\chi^2(n-p)$
3. De plus ces 2 variables aléatoires sont indépendantes.
4. Donc  $\frac{a^*(\hat{\beta} - \beta)}{\hat{\sigma}\sqrt{a^*(M^*M)^{-1}a}} \sim T(n-p)$
5. De sorte que si on pose  $\Phi_{n-p}(z) = P(T(n-p) > z)$  et si  $z_{\alpha/2}(n-p)$  est déterminé par

$$\Phi_{n-p}(z_{\alpha/2}(n-p)) = \alpha/2$$

$$[a^*\hat{\beta} - z_{\alpha/2}(n-p)\sqrt{a^*(M^*M)^{-1}a}\hat{\sigma}, a^*\hat{\beta} + z_{\alpha/2}(n-p)\sqrt{a^*(M^*M)^{-1}a}\hat{\sigma}]$$

est un intervalle de confiance pour la quantité  $a^*\beta$ , au niveau d'erreur  $\alpha$ .

**Estimation de  $\sigma^2$**

En utilisant le fait que  $\hat{\sigma}^2 \sim \frac{\sigma^2}{n-p} \chi^2(n-p)$ , et la définition de  $P(\chi^2(k) > c_\alpha(k)) = \alpha$ , on vérifie facilement que

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

est un intervalle de confiance pour la variance au niveau d'erreur  $\alpha$ .

**5.6 Théorème de Gauss Markov et Moindres Carrés pondérés.**

Considérons le modèle suivant :

$$Y = M\beta + \mathcal{E}$$

où  $\mathcal{E}$  est un vecteur gaussien centré, de matrice de covariance  $\sigma^2 G$ .  $G$  est une matrice symétrique définie positive, connue. Un exemple est la matrice

$$G = \begin{pmatrix} v_1 & 0 & \dots & 0 \\ 0 & v_2 & \dots & 0 \\ & \vdots & & \\ 0 & 0 & \dots & v_n \end{pmatrix},$$

qui correspond au fait que les observations sont encore indépendantes mais chaque observation est entachée d'une variance propre (cas hétéroscédastique).

La question que l'on se pose est doit-on, dans ce cas conserver l'estimateur des moindres carrés,  $\hat{\beta} = (M^*M)^{-1}M^*Y$  ?

La question se pose avec d'autant plus d'acuité qu'un autre estimateur peut sembler tout aussi naturel : En effet, on peut assez simplement transformer le modèle (5.6) en modèle linéaire ordinaire  $Z = M'\beta + \varepsilon$  : En posant  $G = BB^*$ ,  $Z = B^{-1}Y$ ,  $M' = B^{-1}M$ ,  $\varepsilon = B^{-1}\mathcal{E}$ . Dans ce nouveau modèle, on peut calculer l'estimateur usuel des moindres carrés (on remarque en particulier que du fait que  $G$  est symétrique, définie positive,  $B$  est inversible) :

$$\tilde{\beta} = (M'^*M')^{-1}M'^*Z = (M^*G^{-1}M)^{-1}M^*B^{-1*}B^{-1}Y = (M^*G^{-1}M)^{-1}M^*G^{-1}Y.$$

Remarques :

1. Remarquons que par définition, cet estimateur rend minimale la quantité :

$$\|B^{-1}Y - B^{-1}M\beta\|^2 = (Y - M\beta)^*G^{-1}(Y - M\beta)$$

qui représente la norme du vecteur  $Y - M\beta$ , dans la norme  $G^{-1}$ , d'où le nom donné à cet estimateur de moindres carrés pondérés.

Si on considère le cas particulier où  $G$  est diagonale, ce nouvel estimateur conduit à minimiser l'expression

$$\sum_{i=1}^n \frac{1}{v_i^2} (Y_i - (M\beta)_i)^2$$

qui tient compte de la crédibilité de chaque observation en raison inverse de sa variance.

2.  $\text{Var}(a^*\tilde{\beta}a) = a^*(M^*G^{-1}M)^{-1}a$ .
3. Une autre façon d'énoncer la remarque 1 est d'observer que

$$P_V^G = M(M^*G^{-1}M)^{-1}M^*G^{-1}$$

est la matrice associée à l'opérateur de projection dans  $V$ , défini avec la métrique  $G^{-1}$ . Rappelons que si  $A$  est une matrice symétrique définie positive de  $\mathbf{R}^n$ ,  $x^*Ay$  définit un produit scalaire sur  $\mathbf{R}^n$  et on peut donc considérer la métrique associée.

Remarquons que dans ce cas les relations matricielles  $P_V = P_V^*$ ,  $P_V^2 = P_V$ ,  $I_n = P_V + P_{V^\perp}$  valides en métrique euclidienne doivent être remplacées par

$$P_V^G = G(P_V^G)^*G^{-1}, (P_V^G)^2 = P_V^G, I_n = P_V^G + P_{V^\perp, G}^G. \quad (5.2)$$

où  $V^{\perp, G}$  désigne le supplémentaire orthogonal de  $V$ , pour le produit scalaire  $G^{-1}$ . Ces relations se démontrent à partir des relations classiques en observant que

$$\|x\|_{G^{-1}}^2 = x^*B^{-1*}B^{-1}x = \|B^{-1}x\|_{I_n}^2.$$

On en déduit facilement que

$$\begin{aligned} P_V^G &= BP_{B^{-1}V}B^{-1}, V^{\perp, G} = B(B^{-1}V)^\perp \\ P_{V^\perp, G}^G &= BP_{(B^{-1}V)^\perp}B^{-1} \end{aligned}$$

△

Nous allons montrer que cet estimateur possède en fait des propriétés d'optimalité très intéressantes :

Considérons une classe particulière d'estimateurs, et notons que tous les estimateurs de ce modèle, que nous avons considéré jusqu'à présent font partie de cette classe.

**Définition 30** *L'estimateur  $\bar{\beta}$  est dit linéaire s'il existe une matrice  $A$  telle que  $\bar{\beta} = AY$ .*

**Théorème 5** *Considérons le modèle  $Y = M\beta + \mathcal{E}$  où  $\mathcal{E}$  est un vecteur aléatoire centré, de matrice de covariance  $\sigma^2G$ .  $G$  est une matrice symétrique définie positive, connue. Si  $\bar{\beta}$  est un estimateur linéaire, tel que*

$$\mathbf{E}_\beta \bar{\beta} - \beta = 0, \quad \forall \beta \in \mathbf{R}^p,$$

*Alors, il existe  $R$  matrice symétrique positive de  $\mathbf{R}^p$ , telle que  $\text{Var}(\bar{\beta}) = \text{Var}(\tilde{\beta}) + R$ .*

*Remarque :* La signification de ce théorème, est que  $\forall a \in \mathbf{R}^p$ ,  $\text{Var}(a^*\bar{\beta}) \geq \text{Var}(a^*\tilde{\beta})$ . Or cette inégalité est très importante, en particulier si le vecteur  $\mathcal{E}$  est gaussien et que l'on veut construire un intervalle de confiance. En suivant la démarche du paragraphe précédent, on montre très facilement que dans le cas  $\sigma$  connu, cet intervalle est

$$[a^*\bar{\beta} - z_{\alpha/2}\sqrt{\text{Var}(a^*\bar{\beta})}\sigma, a^*\bar{\beta} + z_{\alpha/2}\sqrt{\text{Var}(a^*\bar{\beta})}\sigma]$$

si on utilise  $\bar{\beta}$  et

$$[a^*\hat{\beta} - z_{\alpha/2}\sqrt{\text{Var}(a^*\hat{\beta})}\sigma, a^*\hat{\beta} + z_{\alpha/2}\sqrt{\text{Var}(a^*\hat{\beta})}\sigma]$$

si on utilise  $\hat{\beta}$ . Il est clair qu'on a intérêt à prendre la seconde solution puisque la longueur de l'intervalle est plus petite.

△

*Preuve :*

1. Remarquons d'abord que la condition  $\mathbf{E}_\beta \bar{\beta} - \beta = 0$ ,  $\forall \beta \in \mathbf{R}^p$ , se traduit encore par  $(AM - I_n)\beta = 0$ ,  $\forall \beta \in \mathbf{R}^p$ , c'est à dire  $AM = I_n$ .
2. Par ailleurs,  $\text{Var}(\bar{\beta}) = AGA^*$ . Mais on a  $I_n = P_V^G + P_{V^\perp, G}^G$ , en utilisant (5.2). On en déduit :

$$\begin{aligned}
 \text{Var}(\bar{\beta}) &= A(P_V^G + P_{V^\perp, G}^G)GA^* \\
 &= AM(M^*G^{-1}M)^{-1}M^*G^{-1}GA^* + AP_{V^\perp, G}^GGA^* \\
 &= AM(M^*G^{-1}M)^{-1}M^*A^* + R \\
 &= \text{Var}(\tilde{\beta}) + R
 \end{aligned}$$

3. On finit la démonstration en remarquant que

$$R = AP_{V^\perp, G}^GGA^* = ABP_{B^{-1}V^\perp}B^{-1}BB^*A^* = ABP_{B^{-1}V^\perp}B^*A^*$$

Cette quantité est bien symétrique et positive par les propriétés de la projection en métrique euclidienne.

△



## Chapitre 6

# METHODES BAYESIENNES.

### 6.1 Méthodes Bayésiennes, Introduction.

À la différence des méthodes d'estimation que nous avons introduites précédemment, celles-ci nécessitent d'introduire brièvement la théorie de la décision et des éléments de comparaisons des estimateurs.

#### 6.1.1 Comparaison des estimateurs, théorie de la décision.

Étant donné une expérience dominée  $\mathcal{E}$ , on suppose que l'on désire estimer une quantité  $q(\Theta)$  prenant ses valeurs dans un sous ensemble  $\mathcal{D}$  de  $\mathbf{R}^k$ .

On va 'quantifier' le risque que l'on prend en choisissant une méthode d'estimation. Pour cela, on introduit une fonction de perte :

$$l : \mathcal{D} \times \mathcal{D} \mapsto \mathbf{R}^+$$

qui quantifie par  $l(t, q(\theta))$  la perte que l'on fait quand on estime par  $t$  la quantité  $q(\theta)$ .

On fera généralement les hypothèses suivantes sur  $l$  :

- $l(t, t) = 0$ .
- Il existe une fonction  $\rho$ , positive et convexe telle que

$$l(t, q(\theta)) = \rho(\|t - q(\theta)\|)$$

Les exemples les plus fréquents correspondent à  $l(t, q(\theta)) = \|t - q(\theta)\|$  (perte  $\mathbf{L}_1$ ),  $l(t, q(\theta)) = \|t - q(\theta)\|^2$  (perte quadratique), ou encore la perte suivante (souvent employée si  $\Theta$  est un ensemble fini) :  $l(t, q(\theta)) = 1$  si  $t \neq q(\theta)$ .

À cette fonction de perte et un estimateur  $T$  de  $q(\theta)$ , on associe la fonction de risque :

$$\theta \in \Theta \mapsto R(T, \theta) = \mathbf{E}_\theta l(T, q(\theta)).$$

Le problème que l'on peut légitimement se poser alors, consiste à trouver une procédure d'estimation  $T^*$  qui rende ce risque uniformément minimum. Le théorème suivant montre que cette recherche est impossible, dans beaucoup de cas :

**Théorème 1** *Si le modèle est dominé, si la fonction de perte  $l$  est positive, nulle seulement sur la diagonale, alors si  $q(\theta)$  n'est pas dénombrable, il n'existe pas d'estimateur uniformément de risque minimum.*

*Preuve :* Procédons par l'absurde, supposons qu'il existe  $T^*$ , uniformément optimal i.e. vérifiant :

$$\forall \theta \in \Theta, \mathbf{E}_\theta l(T, q(\theta)) \geq \mathbf{E}_\theta l(T^*, q(\theta))$$

Prenons  $\theta_0$  arbitraire dans  $\Theta$  et comparons les performances de  $T^*$  et de l'estimateur  $T_{\theta_0}$  constamment égal à  $\theta_0$ . Comme  $R(T_{\theta_0}, \theta_0) = 0$ , pour conserver sa souveraineté,  $T^*$  doit vérifier aussi  $R(T^*, \theta_0) = \mathbf{E}_{\theta_0} l(T^*, q(\theta_0)) = 0$ , mais comme  $l$  n'est nulle que sur la diagonale, ceci implique nécessairement

$$T^* = q(\theta_0), P_{\theta_0} p.s.$$

Ou encore, que la loi image  $Q_{\theta_0}$  de  $P_{\theta_0}$  par  $T^*$  est une masse de Dirac en  $q(\theta_0)$ .

Par ailleurs, on sait que si la famille  $\{P_\theta, \theta \in \Theta\}$  est dominée par une mesure  $\mu$  alors la mesure image de  $\mu$  par  $T^*$  domine la famille  $\{Q_\theta, \theta \in \Theta\}$ . Mais ceci nous mène à une contradiction puisqu'on a montré qu'un ensemble de masses de Dirac ne pouvait être dominé que s'il était dénombrable.  $\triangle$

*Exercice :* Par ailleurs, cette preuve est illustrée par l'exemple simple suivant : On observe  $X$  qui suit une loi binomiale  $B(n, p), p \in (0, 1)$ . On veut estimer  $p$ . On a déjà rencontré l'estimateur  $T = \frac{X}{n}$ , mais on pourrait penser aussi à la famille d'estimateurs  $T_{b,r} = \frac{X+r}{n+b}$ . Représenter graphiquement les risques quadratiques de ces estimateurs. Existe-t-il parmi ces estimateurs un estimateur uniformément de risque minimal, un estimateur de risque constant ?  $\triangle$

Pour contourner cette difficulté, on a généralement recours à plusieurs types de stratégies :

- Soit on restreint la classe des estimateurs. Par exemple, on peut ne considérer que des estimateurs 'sans biais', ce qui a l'avantage d'exclure des estimateurs absurdes comme  $T_{\theta_0}$ . Néanmoins, cette stratégie peut aussi poser des problèmes : la classe des estimateurs sans biais peut être très petite, elle peut même être vide.
- On peut, aussi sans restreindre la classe des estimateurs, décider d'optimiser

$$\sup_{\theta \in \Theta} R(T, \theta).$$

L'estimateur optimum est appelé 'minimax', parce qu'il minimise le risque maximum. Ce point de vue qui est beaucoup employé dans les situations non paramétriques est considéré par ses détracteurs comme trop pessimiste...

- Enfin on peut, toujours sans restreindre la classe des estimateurs, décider d'optimiser un 'risque moyen

$$\int_{\Theta} R(T, \theta) d\nu(\theta).$$

C'est le point de vue des méthodes bayésiennes.

### 6.1.2 Loi a priori, Contexte bayésien.

La différence fondamentale du contexte bayésien avec le contexte classique réside dans l'introduction d'une loi de probabilité *a priori*  $\nu$  sur l'ensemble des paramètres. Cela nécessite au préalable de munir  $\Theta$  d'une tribu  $\mathcal{T}$ . La loi  $\nu$  reflète alors, ce qu'on est sensé savoir du paramètre, *avant* l'expérience.

Ceci n'est pas sans conséquence sur notre modèle, puisque, de ce fait,  $\theta$  est une variable aléatoire, et donc  $P_\theta$ , représente maintenant la loi de l'observation  $X$ , conditionnellement à  $\theta$ .

On appelle alors *loi conjointe* la loi du vecteur  $(X, \theta)$  et loi *a posteriori* la loi de  $\theta$  conditionnelle à l'observation  $X$  ( $\theta|X$ ) qui reflète alors, ce que l'on sait sur le paramètre *après* l'expérience.

Étant donné une fonction de perte définie comme au paragraphe précédent, et un estimateur  $T$  de la quantité  $q(\theta)$ , on définit alors le risque *bayésien* de  $T$ ,

$$R(T, \nu) = \int_{\Theta} R(T, \theta) d\nu(\theta).$$

On a alors la définition suivante :

**Définition 31** Dans le cadre précédent, un estimateur  $T^*$  est dit bayésien associé à la fonction de perte  $l$  et à la mesure a priori  $\nu$ , s'il vérifie :

$$R(T^*, \nu) \leq R(T, \nu)$$

pour tout estimateur  $T$ .

## 6.2 Calcul de loi a posteriori, Exemples

Notons maintenant  $p(x|\theta)$  une densité de  $P_\theta$  par rapport à la mesure dominante  $\mu$ . (Nous supposons toujours le modèle dominé.) Notons que le changement de notation correspond à la nouvelle interprétation dans le cadre bayésien de la loi  $P_\theta$ .

Pour faciliter les calculs, nous considérerons une mesure  $m$  sur  $(\Theta, \mathcal{T})$  qui domine  $\nu$ , et nous noterons  $n(\theta)$ , une densité de  $\nu$  par rapport à  $m$ .

Il est alors facile de vérifier que la loi conjointe de  $(X, \theta)$  sur  $(\mathcal{X} \times \Theta)$ ,  $(\mathcal{A} \otimes \mathcal{T})$  est dominée par la mesure produit  $\mu \otimes m$  par rapport à laquelle elle admet la densité :

$$\pi(x, \theta) = p(x|\theta)n(\theta).$$

Par le théorème de Bayes, la loi a posteriori sur  $\Theta, \mathcal{T}$  est aussi dominée par  $m$ , et admet la densité :

$$p(\theta|x) = \frac{p(x|\theta)n(\theta)}{\int_{\Theta} p(x|\theta)n(\theta)dm(\theta)}$$

*Exemple :* Prenons à nouveau, le cas du modèle binomial où le paramètre inconnu est  $\theta = p \in \Theta = [0, 1]$ . Le modèle est dominé par la mesure  $\mu = \sum_{k=0}^n \delta_k$  et

$$p(x|\theta) = C_n^k \theta^x (1 - \theta)^{n-x}$$

Supposons que l'on choisisse la loi a priori de la façon suivante : (Ce choix sera discuté ultérieurement.) On prend pour  $\nu$  une loi  $Beta(r, s)$ .

On rappelle que pour des paramètres  $r$  et  $s$  strictement positifs, on appelle loi  $Beta(r, s)$ , la loi dont la densité par rapport à la mesure ( $m$ , ici) de Lebesgue sur  $[0, 1]$  est donnée par

$$n(\theta) = c(r, s)\theta^{r-1}(1-\theta)^{s-1}.$$

On rappelle que  $c(r, s) = \left[ \int_{[0,1]} \theta^{r-1}(1-\theta)^{s-1} d\theta \right]^{-1}$ , que la moyenne de cette loi est  $\frac{r}{r+s}$ , et sa variance est  $\frac{rs}{(r+s)(r+s+1)}$ .

La loi conjointe admet alors une densité par rapport à  $\mu \otimes m$  donnée par :

$$\pi(x, \theta) = c(r, s) C_n^k \theta^{x+r-1} (1-\theta)^{n-x+s-1}.$$

La loi a posteriori admet par rapport à  $m$  la densité :

$$p(\theta|x) = \frac{\theta^{x+r-1}(1-\theta)^{n-x+s-1}}{\int_{[0,1]} \theta^{x+r-1}(1-\theta)^{n-x+s-1} d\theta} = c(r+x, n-x+s) \theta^{x+r-1} (1-\theta)^{n-x+s-1}$$

C'est donc une loi  $Beta(r+x, s+n-x)$ . (Ne pas perdre de vue que  $x$  est notre observation, c'est donc une quantité aléatoire.)

Ceci nous permet d'interpréter les paramètres  $r, s$  de la loi a priori. En effet, en observant comment s'opère la modification de notre connaissance sur le paramètre avant et après observation, on remarque que  $r$  et  $x$  jouent des rôles analogues, de même pour  $r+s$  et  $n$ . On peut donc interpréter la loi a priori comme une observation préalable à l'expérience, portant sur  $r+s$  observations (au sens où une binomiale  $B(n, \theta)$  peut toujours être considérée comme la somme de  $n$  variables de Bernoulli indépendantes), et au cours de laquelle l'observation aurait été  $x' = r$ .

Le fait que les lois a priori et a posteriori se retrouvent dans la même famille de lois n'est pas un hasard. On dit alors que cette famille de lois est conjuguée au modèle. Nous verrons d'autres exemples de ce phénomène.

△

## 6.3 Calcul de l'estimateur bayésien.

### 6.3.1 Perte quadratique ou de type $L_1$ .

Nous nous plaçons maintenant dans le cas suivant :  $\Theta \subset \mathbf{R}$ ,  $q(\theta) = \theta$ . Nous allons démontrer les théorèmes suivants :

**Théorème 2** avec les notations précédentes, si la fonction de perte est :

$$l(t, \theta) = (t - \theta)^2$$

si le modèle et la loi a priori sont choisis de sorte que :

$$\int_{\Theta} \theta^2 p(\theta|x) dm(\theta) < +\infty, \forall x \in \mathcal{X}, \mu - p.s.$$

alors l'estimateur bayésien du problème est donné par

$$T^*(x) = \int_{\Theta} \theta p(\theta|x) dm(\theta)$$

**Théorème 3** avec les notations précédentes, si la fonction de perte est :

$$l(t, \theta) = |t - \theta|$$

si le modèle et la loi a priori sont choisis de sorte que :  $\forall x$  dans  $\mathcal{X}$ ,  $\mu - p.s.$ , il existe  $\tau(x)$  vérifiant

$$\int_{\theta \leq \tau(x)} p(\theta|x) dm(\theta) = \int_{\theta \geq \tau(x)} p(\theta|x) dm(\theta) = 1/2.$$

( $\tau(x)$  est unique médiane de la loi a posteriori.) alors l'estimateur bayésien du problème est donné par

$$T^*(x) = \tau(x)$$

Les deux théorèmes sont une conséquence des lemmes suivants.

**Lemme 5** Avec les notations précédentes, pour que  $T^*$  soit un estimateur bayésien associé à la fonction de perte  $l$ , il suffit que, pour tout  $x$  dans  $\mathcal{X}$ ,  $\mu - p.s.$ ,  $T^*(x)$  minimise la fonction :

$$r \in \mathbf{R} \mapsto \int_{\Theta} l(r, \theta) p(\theta|x) dm(\theta)$$

Démonstration du lemme 5 :

Définissons la marginale en  $X$ , de densité par rapport à la mesure  $\mu(x)$ ,

$$p(x) = \int_{\Theta} \pi(x, \theta) dm(\theta).$$

Il suffit de remarquer qu'on cherche à minimiser (en  $T(x)$ ) la quantité suivante, que l'on transforme en utilisant le théorème de Fubini :

$$\begin{aligned} \int_{\Theta} [R(T, \theta)] d\nu(\theta) &= \int_{\Theta} \left[ \int_{\mathcal{X}} l(T(x), \theta) p(x|\theta) d\mu(x) \right] n(\theta) dm(\theta) \\ &= \int_{\Theta} \int_{\mathcal{X}} l(T(x), \theta) \pi(x, \theta) d\mu(x) dm(\theta) \\ &= \int_{\Theta} \int_{\mathcal{X}} l(T(x), \theta) p(\theta|x) p(x) d\mu(x) dm(\theta) \\ &= \int_{\mathcal{X}} \left\{ \int_{\Theta} l(T(x), \theta) p(\theta|x) dm(\theta) \right\} p(x) d\mu(x) \end{aligned}$$

On voit alors que si on minimise la quantité entre parenthèses pour tout  $x$  dans  $\mathcal{X}$ ,  $\mu - p.s.$ , on minimisera à coup sûr l'intégrale. ■

**Lemme 6** Si  $Z$  est une variable aléatoire réelle, telle que  $\mathbf{E}Z^2 < \infty$ , alors la fonction :  
 $r \in \mathbf{R} \mapsto \mathbf{E}(Z - r)^2$  admet un unique minimum en  $r = \mathbf{E}Z$ .

Démonstration du lemme 6 : On remarque simplement que :

$$\mathbf{E}(Z - r)^2 = \mathbf{E}(Z - \mathbf{E}Z)^2 + (r - \mathbf{E}Z)^2$$

**Lemme 7** Si  $Z$  est une variable aléatoire réelle, telle qu'il existe  $\tau$ ,  $P(Z \leq \tau) = P(Z \geq \tau) = 1/2$  alors la fonction :  
 $r \in \mathbf{R} \mapsto \phi(r) = \mathbf{E}|Z - r|$  admet un unique minimum pour  $r = \tau$ .

Démonstration du lemme 7

1. Remarquons d'abord que  $\phi$  est une fonction convexe : Pour tout  $\lambda \in [0, 1]$ ,

$$\begin{aligned} \phi(\lambda r_1 + (1 - \lambda)r_2) &= \mathbf{E}|\lambda(Z - r_1) + (1 - \lambda)(Z - r_2)| \\ &\leq \lambda\phi(r_1) + (1 - \lambda)\phi(r_2) \end{aligned}$$

2. Par ailleurs  $\phi(r) \geq |r - \mathbf{E}Z|$  donc  $\phi$  tend vers l'infini quand  $|r|$  tend vers l'infini.  
 3. Nous pouvons donc en conclure que  $\phi$  admet en tout point une dérivée à gauche et une dérivée à droite et un minimum en un point  $r_0$  vérifiant :  $\phi'((r_0)_-) \leq 0$ ,  $\phi'((r_0)_+) \geq 0$   
 4. On a :

$$\begin{aligned} \phi(r) &= - \int_{x \leq r} (x - r)dP(x) + \int_{x \geq r} (x - r)dP(x) \\ &= \mathbf{E}Z - r - 2 \int_{x \leq r} (x - r)dP(x) \end{aligned}$$

5. On a en utilisant Fubini :

$$\int_{x \leq r} F(x)dx = \int_{-\infty}^{\infty} I\{x \leq r\} [\int_{-\infty}^{\infty} I\{z \leq x\} dP(z)] dx = \int_{\mathbf{R}^2} I\{z \leq x \leq r\} dx dP(z) = \int_{\mathbf{R}} I\{z \leq r\} (z - r) dP(z)$$

6. On déduit de 4. et 5. que :

$$\phi(r) = \mathbf{E}Z - r + 2 \int_{x \leq r} F(x)dx$$

7. Pour  $h > 0$ , on peut donc écrire :

$$\frac{\phi(r+h) - \phi(r)}{h} = -1 + 2 \frac{1}{h} \int_r^{r+h} F(x)dx \xrightarrow{h \rightarrow 0} -1 + 2F(r_+)$$

8. le lemme s'obtient en faisant un calcul identique pour  $h < 0$  et en utilisant 3.

*Exemples*

1. Reprenons l'exemple du modèle binomial, doté d'une loi a priori de type  $Beta(r, s)$ . On a vu que la loi a posteriori, étant donné une observation  $x \in \{0, \dots, n\}$  était une loi  $Beta(r + x, s + n - x)$ . On peut donc appliquer, par exemple le théorème 2. On obtient alors que l'estimateur bayésien est

$$T^*(x) = \frac{r + x}{n + r + s}.$$

Nous retrouvons la famille d'estimateurs considérée dans le premier paragraphe de ce chapitre. On retrouve aussi les rôles respectifs joués par les paramètres de la loi a priori.

2. Supposons que l'on observe un  $n$ -échantillon de variables aléatoires gaussiennes  $N(\theta, 1)$ . On se propose d'estimer  $q(\theta) = \theta$ .

$t$  et  $v^2$  étant des paramètres arbitrairement fixés, choisissons comme loi a priori sur  $\theta$  une loi normale  $N(t, v^2)$ . On peut alors prendre pour  $\mu$  la mesure de Lebesgue sur  $\mathbf{R}^n$ , et pour  $m$ , la mesure de Lebesgue sur  $\mathbf{R}$ . On a alors, pour  $x = (x_1, \dots, x_n)$

$$\begin{aligned} p(x_1, \dots, x_n | \theta) &= \frac{1}{(2\pi)^{\frac{n}{2}}} \exp \frac{-1}{2} \sum_{i=1}^n (x_i - \theta)^2 \\ \pi(x_1, \dots, x_n, \theta) &= \frac{1}{v(2\pi)^{\frac{n+1}{2}}} \exp \frac{-1}{2} \left[ \sum_{i=1}^n (x_i - \theta)^2 + \frac{(\theta - t)^2}{v^2} \right] \\ p(\theta | x_1, \dots, x_n) &= C(x_1, \dots, x_n) \exp \frac{-1}{2} \left[ \sum_{i=1}^n (x_i - \theta)^2 + \frac{(\theta - t)^2}{v^2} \right] \\ &= C'(x_1, \dots, x_n) \exp \frac{-1}{2} \left[ \left( n + \frac{1}{v^2} \right) \left( \theta - \frac{\sum_{i=1}^n x_i + \frac{t}{v^2}}{n + \frac{1}{v^2}} \right)^2 \right] \end{aligned}$$

On déduit de cette dernière écriture que la loi a posteriori est une normale

$$N\left(\frac{\sum_{i=1}^n x_i + \frac{t}{v^2}}{n + \frac{1}{v^2}}, \frac{1}{n + \frac{1}{v^2}}\right)$$

On peut alors facilement appliquer les théorèmes 2 et 3. On obtient pour les deux fonctions de perte le même estimateur :

$$T^*(x_1, \dots, x_n) = \frac{\sum_{i=1}^n x_i + \frac{t}{v^2}}{n + \frac{1}{v^2}}.$$

On voit bien tant sur la loi a posteriori que sur l'estimateur la façon dont on peut interpréter les différents paramètres de la loi a priori : Elle s'interprète à nouveau comme une observation préalable ayant portée sur un  $n' \approx \frac{1}{v^2}$ -échantillon ( $n$  joue le même rôle que  $\frac{1}{v^2}$ ), et donnant des observations dont la moyenne est  $t$  ( $t$  joue le même rôle que  $\frac{\sum_{i=1}^n x_i}{n}$ ).

3. (Exercice) Reprendre le modèle du  $n$ -échantillon gaussien. Supposons maintenant qu'il s'agit de gaussienne  $N(\theta, \sigma^2)$  où  $\sigma^2$  aussi est inconnu. Quelle famille de loi a priori doit-on choisir, pour que la loi a posteriori reste dans cette famille ?

### 6.3.2 Problème de classification.

Étudions maintenant le problème suivant très important en pratique : On observe le vecteur aléatoire de  $\mathbf{R}^k$ ,  $Y$ . On sait que la loi du vecteur  $Y$  se trouve nécessairement parmi les lois  $N(\beta_1, \Gamma), \dots, N(\beta_l, \Gamma)$ .  $\beta_1, \dots, \beta_l$  sont des vecteurs connus (et différents) de  $\mathbf{R}^k$ ,  $\Gamma$  est une matrice de covariance de dimension  $k \times k$ , connue et définie positive.

Notre problème est donc simplement de choisir entre les  $\beta_i$ .

Nous nous placer en contexte bayésien et mettre une loi a priori sur notre ensemble de paramètres :

$$\nu_i = \nu\{\beta = \beta_i\}.$$

Nous allons considérer avec un intérêt particulier le cas où  $\nu_i = \frac{1}{l}$ . Il correspond au fait de ne vouloir privilégier aucune des hypothèses.

Nous prenons pour perte la fonction :

$$l(\beta, \beta_i) = 1_{\beta \neq \beta_i}.$$

Pour trouver l'estimateur  $\beta^*(Y) \in \{\beta_1, \dots, \beta_l\}$ , nous allons donc minimiser le risque bayésien du problème :

$$\begin{aligned} \sum_{i=1}^l \mathbf{E}_{\beta_i} l(\beta^*(Y), \beta_i) \nu_i &= \sum_{i=1}^l \mathbf{E}_{\beta_i} 1_{\beta^*(Y) \neq \beta_i} \nu_i \\ &= \sum_{i=1}^l \int_{\mathbf{R}^k} 1_{\beta^*(Y) \neq \beta_i} p(y, \beta_i) dy \nu_i \\ &= \int_{\mathbf{R}^k} \left[ \sum_{i=1}^l 1_{\beta^*(Y) \neq \beta_i} p(y, \beta_i) \nu_i \right] dy \end{aligned}$$

Il est clair, sur cette dernière expression que si on emploie la stratégie suivante :

$$\beta^* = \beta_{i^*}$$

avec  $i^* = \text{Arg sup}_i p(y, \beta_i) \nu_i$ , on minimisera certainement le risque bayésien.

Il est en particulier intéressant de considérer le cas  $\nu_i = \frac{1}{l}$ . Un calcul simple montre que dans ce cas, on a

$$i^* = \text{Arg inf}_i (y - \beta_i)^* \Gamma^{-1} (y - \beta_i)$$

Ce qui correspond à choisir le vecteur des moyennes qui est le plus près de l'observation  $y$  au sens de la forme quadratique associée à l'inverse de la covariance.

Exercice : Etudier le cas où  $l = 2$  et comparer le résultat trouvé au théorème de Neymann-Pearson.

# Chapitre 7

## Tests

### 7.1 Introduction

La théorie des tests est une autre partie importante de la statistique. La problématique est la suivante.

On se donne un modèle  $\mathcal{E} = (X, \mathcal{X}, \mathcal{F}, P_\theta, \theta \in \Theta)$ . On se donne une partition de  $\Theta$  en deux ensembles (non vides)  $\Theta_0$  et  $\Theta_1$ . Le but du jeu est alors de décider si  $\theta$  appartient à  $\Theta_0$  ou  $\Theta_1$ .

**Définition 32** Dans le contexte exposé ci-dessus une variable aléatoire  $\phi(X)$  à valeurs dans  $\{0, 1\}$  est appelée **test**. La procédure de décision associée consiste à décider  $\Theta_0$  si  $\phi(x) = 0$  et  $\Theta_1$  sinon.

Une variable aléatoire à valeurs dans  $[0, 1]$  est appelée **test randomisé**. La procédure de décision associée consiste si  $\phi(x) = \gamma$  à tirer au sort à l'aide d'une variable de Bernoulli de paramètre  $\gamma$  (*i.e.* ;  $P\{1\} = 1 - P\{0\} = \gamma$ ) indépendante de l'expérience et à décider  $\Theta_0$  ou  $\Theta_1$  en fonction du tirage obtenu.

#### Notation :

On note généralement :

$$\begin{aligned} \mathcal{H}_0, & \quad \text{l'hypothèse 'nulle' : } \quad \{\theta \in \Theta_0\} \\ \mathcal{H}_1, & \quad \text{'l'alternative' : } \quad \{\theta \in \Theta_1\} \end{aligned}$$

Quand on fait un test, il y a deux façon de se tromper, déclarer  $\mathcal{H}_1$  alors que  $\mathcal{H}_0$  est vrai ou l'inverse. Ceci conduit aux deux définitions suivantes :

**Définition 33** Etant donnée l'expérience  $\mathcal{E}$  et le problème de test associé à la partition  $\Theta_0, \Theta_1$ ,  $\alpha \in [0, 1]$ , on dit que le test  $\phi(X)$  est de **niveau**  $\alpha$  ssi

$$\sup_{\theta \in \Theta_0} \mathbf{E}_\theta \phi(X) \leq \alpha$$

**Définition 34** *Etant donnée l'expérience  $\mathcal{E}$  et le problème de test associé à la partition  $\Theta_0, \Theta_1$ ,  $\alpha \in [0, 1]$ , on appelle erreur de deuxième espèce (resp. puissance) la fonction*

$$\theta \in \Theta_1 \mapsto \mathbf{E}_\theta(1 - \phi(X)) \quad (\text{resp. } \mathbf{E}_\theta\phi(X))$$

Nous allons nous attacher à construire des tests de niveau  $\alpha$  et tels que l'erreur de deuxième espèce soit raisonnable ou éventuellement la plus petite possible.

## 7.2 Test d'une hypothèse simple, lien avec la Théorie de l'estimation.

Considérons le cas où  $\Theta$  est inclus dans  $\mathbf{R}$ ,  $\Theta_0 = \{\theta_0\}$ ,  $\Theta_1 = \Theta \setminus \{\theta_0\}$ , et où on dispose d'un intervalle de confiance au niveau  $\alpha$  pour estimer  $\theta : [R(X), T(X)]$ . Une simple réécriture de la définition précédente nous montre que le test non randomisé suivant

$$\phi(X) = 0, \text{ si } \theta_0 \in [R(X), T(X)], \quad 1, \quad \text{sinon}$$

est de niveau  $\alpha$ .

Dans ce cas, il est facile de calculer l'erreur de deuxième espèce :

$$\theta \in \Theta_1 \mapsto P_\theta(\theta_0 \in [R(X), T(X)])$$

## 7.3 Test d'une sous hypothèse linéaire.

Plaçons-nous maintenant, dans le cadre d'un modèle linéaire, dont la matrice exogène  $M$  est de rang  $p$ . On se donne  $C$ , une matrice fixée de dimension  $l \times p$ , avec  $l < p$ , on suppose que le rang de  $C$  est  $l$  et on se propose de tester l'hypothèse

$$\mathcal{H}_0 = \{\beta \in \mathbf{R}^p / C\beta = 0\}.$$

*Exemples :*

1. Si  $l = 1$ , on se ramène à tester la nullité d'une forme linéaire. On retrouve donc l'étude du paragraphe précédent. En effet, dans le chapitre sur les méthodes de contrastes, nous avons construit des intervalles de confiance pour les quantités  $a^*\beta$ .
2. Prenons par exemple le cas où  $Y_i$  est la mesure d'un taux de pollution, que l'on cherche à expliquer par différentes variables :  $M^1$  quantité de précipitations,  $M^2$  vitesse du vent,  $M^3$  température,  $M^4$  nombre d'usines, à travers le modèle suivant :

$$Y_i = \beta_1 M_i^1 + \beta_2 M_i^2 + \beta_3 M_i^3 + \beta_4 M_i^4 + \varepsilon_i$$

Il est clair que plus le modèle contient de paramètres, en général, moins il est interprétable. Donc on peut se poser la question de diminuer le nombre de paramètres, par exemple, en testant  $\beta_1 = \beta_3 = 0$ .

$\triangle$

### 7.3.1 Résolution

1. Soit  $V_1$  le sous espace vectoriel de  $V$ ,

$$V_1 = \{M\beta, C\beta = 0\}$$

Comme  $\text{rg}(C) = l, \dim(V_1) = \dim(\ker(C)) = p - l$ .

2. Soit  $W_1$  le supplémentaire orthogonal de  $V_1$  dans  $V$ . On a

$$I_n = P_{V_1} + P_{W_1} + P_{V_\perp},$$

$P_{V_1}, P_{W_1}, P_{V_\perp}$  sont des projecteurs respectivement de rang  $p - l, l, n - p$

3. En appliquant le théorème de Cochran, on a que

(a)  $(\sigma)^{-1}P_{V_1}\varepsilon, (\sigma)^{-1}P_{W_1}\varepsilon, (\sigma)^{-1}P_{V_\perp}\varepsilon$  sont des vecteurs gaussiens, indépendants de lois respectives  $N(0, P_{V_1}), N(0, P_{W_1}), N(0, P_{V_\perp})$ .

(b) D'où,  $(\sigma)^{-1}P_{V_1}Y, (\sigma)^{-1}P_{W_1}Y, (\sigma)^{-1}P_{V_\perp}Y$  sont des vecteurs gaussiens indépendants de lois respectives  $N(P_{V_1}M\beta, P_{V_1}), N(P_{W_1}M\beta, P_{W_1}), N(0, P_{V_\perp})$ .

(c) On en déduit que :

i.  $\|(\sigma)^{-1}P_{V_\perp}Y\|^2 \sim \chi^2(n - p)$ .

ii.  $\|(\sigma)^{-1}P_{V_\perp}Y\|^2$  et  $\|(\sigma)^{-1}P_{W_1}Y\|^2$  sont indépendants.

iii. – Si  $C\beta = 0, P_{W_1}(M\beta) = 0$  et donc  $\|(\sigma)^{-1}P_{W_1}Y\|^2 \sim \chi^2(l)$ .  
– Si  $C\beta \neq 0, \|(\sigma)^{-1}P_{W_1}Y\|^2 \sim \chi^2(l, \|(\sigma)^{-1}P_{W_1}(M\beta)\|^2)$ .

4. On en déduit que sous l'hypothèse  $\mathcal{H}_0$ , la statistique

$$T = \frac{\|P_{W_1}Y\|^2/l}{\|P_{V_\perp}Y\|^2/(n - p)} \sim F(l, n - p).$$

5. D'où, si  $f_\alpha(n_1, n_2)$ , est déterminé par  $P(F(n_1, n_2) > f_\alpha(n_1, n_2)) = \alpha$ , on a

$$1 - \alpha = P\left(T \in [f_{1-\alpha/2}(l, n - p), f_{\alpha/2}(l, n - p)]\right).$$

6. Donc,

- Si la statistique  $T$ , évaluée sur nos données, tombe en dehors de l'intervalle  $[f_{1-\alpha/2}(l, n - p), f_{\alpha/2}(l, n - p)]$ , on rejettera l'hypothèse  $\mathcal{H}_0$ .
- En revanche, si elle tombe dans cet intervalle, on acceptera l'hypothèse.

Cette stratégie fournit un test de niveau  $\alpha$  par construction.

### 7.3.2 Calcul pratique de $T$

On a

$$T = \frac{\|M\hat{\beta} - P_{V_1}Y\|^2/l}{\|Y - M\hat{\beta}\|^2/(n - p)}$$

1. Si

$$C = \begin{pmatrix} 1 & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 & 0 & \dots & 0 \\ & & \vdots & & & & \\ 0 & 0 & \dots & 1 & 0 & \dots & 0 \end{pmatrix},$$

dans ce cas, on cherche à tester  $\beta_1 = \dots = \beta_l = 0$ .

Soit  $\tilde{M} = (M_{l+1}, \dots, M_p)$ , la matrice des  $l - p$  vecteurs colonnes de  $M$ .

Il est facile de montrer que  $P_{V_1} Y = \tilde{M}(\tilde{M}^* \tilde{M})^{-1} \tilde{M}^* Y$ , et  $T$  se calcule aisément en fonction de  $M$  et  $\tilde{M}$ .

2. Dans le cas général, où  $C$  est une matrice quelconque, on commence par compléter  $C$  en une matrice  $C'$   $p \times p$  et inversible, puis on pose  $\eta = C' \beta$ . Le modèle linéaire  $Y = M \beta + \varepsilon$  est équivalent au modèle linéaire suivant, dans lequel on a fait le changement de paramètre  $\mu = C' \beta$ ,  $M' = M C'^{-1}$  :

$$Y = M' \mu + \varepsilon.$$

Dans ce nouveau modèle l'hypothèse à tester est  $\mu_1 = \dots = \mu_l = 0$  et on est ramené au cas précédent.

## 7.4 Test de Neyman- Pearson.

On peut se poser le problème de construire un test qui soit de niveau  $\alpha$  donné, et optimal parmi les tests de niveau  $\alpha$ , c'est à dire tel que l'erreur de deuxième espèce soit la plus petite possible. Ce problème est en général sans solution. Néanmoins, dans le cas où l'on teste une hypothèse simple contre une hypothèse simple (i.e.  $\Theta_0 = \{\theta_0\}$ ,  $\Theta_1 = \{\theta_1\}$ ) le problème a une solution (au moins dans les tests randomisés) donnée par le théorème suivant appelé Théorème de Neyman-Pearson.

**Théorème 6** Soient  $P_0$  et  $P_1$  2 lois de probabilités sur un même espace,  $p_1$  et  $p_0$  leurs densités respectives par rapport à une mesure dominante  $\mu$ . Soit  $\alpha \in (0, 1)$ ,

-  $\exists k \in \mathbf{R}$ ,  $\gamma \in (0, 1)$  /

$$\begin{aligned} \Phi^*(x) &= 1 \text{ si } p_1(x) > k p_0(x) \\ &\quad \gamma \text{ si } p_1(x) = k p_0(x) \\ &= 0 \text{ si } p_1(x) < k p_0(x) \end{aligned}$$

est de niveau  $\alpha$  (en fait,  $\mathbf{E}_0 \Phi^* = \alpha$ )

- Soit  $\Phi$  un test de niveau  $\alpha$ , alors

$$\mathbf{E}_1 \Phi \leq \mathbf{E}_1 \Phi^* \tag{7.1}$$

- Si  $\Phi$  est un test de niveau  $\alpha$ , tel que  $\mathbf{E}_1 \Phi = \mathbf{E}_1 \Phi^*$ , alors  $\Phi$  coïncide avec  $\Phi^*$  sur l'ensemble  $\{x : p_1(x) \neq k p_0(x)\}$  à un ensemble de  $\mu$  mesure nulle près.

*Remarques :*

1. La première assertion est simplement une construction du test optimal. Notons, que la construction ainsi présentée dépend de la mesure dominante  $\mu$  (que l'on peut sans perte de généralité prendre égale à  $P_0 + P_1$ ).

2. La seconde assertion prouve effectivement l'optimalité du test  $\Phi^*$ .
3. La troisième assertion prouve une 'sorte' d'unicité du test optimal.

△

### Démonstration du théorème :

- Considérons la fonction définie sur  $\mathbf{R}$  qui à  $t$  associe

$$G_0(t) = P_0\{x : p_1(x) \leq tp_0(x)\}.$$

Il est facile de voir que c'est une fonction de répartition donc elle est croissante, continue à droite limitée à gauche. De plus quand  $G_0(t) \neq G_0(t_-)$ , alors

$$P_0\{x : p_1(x) = tp_0(x)\} = G_0(t) - G_0(t_-).$$

Posons

$$t_\alpha = \sup\{t / G_0(t) \leq 1 - \alpha\}$$

Alors de deux choses l'une : Soit  $G_0(t_\alpha) = 1 - \alpha$  et dans ce cas si  $k = t_\alpha \gamma = 0$  conviennent. Soit  $G_0([t_\alpha]_-) \leq 1 - \alpha < G_0(t_\alpha)$ . Dans ce cas, il est facile d'observer que

$$P_0\{x : p_1(x) = t_\alpha p_0(x)\} = G_0(t_\alpha) - G_0([t_\alpha]_-) > 0$$

et que  $k = t_\alpha$ ,  $\gamma = \frac{G_0(t_\alpha) - 1 + \alpha}{G_0(t_\alpha) - G_0([t_\alpha]_-)}$  conviennent. Ceci démontre la première assertion.

- Soit  $\phi$  un test de niveau  $\alpha$ . Considérons la fonction qui à  $x$  associe

$$f(x) = [\Phi(x) - \Phi^*(x)][p_1(x) - kp_0(x)]$$

- Si  $p_1(x) > kp_0(x)$ ,  $\Phi^*(x) = 1 \geq \Phi(x)$ , donc  $f(x) \leq 0$ .
- Si  $p_1(x) < kp_0(x)$ ,  $\Phi^*(x) = 0 \leq \Phi(x)$ , de sorte que  $f(x) \leq 0$
- Si  $p_1(x) = kp_0(x)$ ,  $f(x) = 0$

On a donc montré que  $f(x) \leq 0$ ,  $\forall x$ . Il en résulte que :

$$0 \geq \int f(x)d\mu = \mathbf{E}_1[\Phi - \Phi^*] - k\mathbf{E}_0[\Phi - \Phi^*] \geq \mathbf{E}_1[\Phi - \Phi^*]$$

Car  $\mathbf{E}_0[\Phi - \Phi^*] \leq 0$  si  $\Phi$  est de niveau  $\alpha$ .

- Si maintenant  $\Phi$ , de niveau  $\alpha$  vérifie  $\mathbf{E}_1\Phi = \mathbf{E}_1\Phi^*$ , alors,  $f$  est une fonction négative dont l'intégrale est nulle. Ceci implique qu'elle est nulle  $\mu$  presque partout, ce qui entraîne la troisième assertion. ■

**Exemple :** On observe un  $n$ -échantillon de variables gaussiennes  $N(\mu, \sigma^2)$ . On suppose que  $\sigma > 0$  est connu et  $\Theta = \{\mu_0, \mu_1\}$ ,  $\mu_1 > \mu_0$ .

On se propose de tester  $\mathcal{H}_0 : \mu = \mu_0$ . Pour déterminer la forme du test optimal, on procède comme dans la démonstration du théorème précédent.

On considère la fonction qui à  $t \in \mathbf{R}$  associe

$$G_0(t) = P_{\mu_0}(\exp \frac{-(x - \mu_1)^2 + (x - \mu_0)^2}{2\sigma^2} \leq t).$$

On remarque que, pour  $t \leq 0$ ,  $G_0(t) = 0$ , pour  $t > 0$ ,

$$\begin{aligned} G_0(t) &= P_{\mu_0, \sigma^2} \left( x \leq \frac{\mu_0 + \mu_1}{2} + \frac{\log t}{\mu_1 - \mu_0} \right) \\ &= P_{0,1}(x \leq u(t)) \end{aligned}$$

si on pose  $u(t) = \frac{1}{\sigma} \left[ \frac{\mu_1 - \mu_0}{2} + \frac{\log t}{\mu_1 - \mu_0} \right]$ . On voit donc que  $G_0(t)$  est une fonction croissante et continue. Donc pour tout  $\alpha \in [0, 1]$ , il existe  $t_\alpha$  tel que  $G_0(t_\alpha) = \alpha$ . Par ailleurs, il est clair que le test optimal est unique, non randomisé et s'écrit alors

$$\begin{aligned} \Phi^*(x) &= 1_{\{x \leq u(t_\alpha)\}} \\ \text{avec } \int_{u(t_\alpha)} &\frac{\exp -\frac{u^2}{2}}{\sqrt{2\pi}} du = \alpha \end{aligned}$$

Les conséquences de cette construction sont

- Que pour construire le test optimal, on n'a pas besoin de calculer  $t_\alpha$ , mais uniquement  $u(t_\alpha)$  que l'on trouve facilement dans les tables.
- Le test ne dépend pas de  $\mu_1$  (à condition que  $\mu_1$  reste plus grand que  $\mu_0$ ). Le test  $\Phi^*$  est donc optimal, non seulement pour tester  $\mu = \mu_0$  contre  $\mu = \mu_1$ , mais aussi  $\mu = \mu_0$  contre  $\mu > \mu_0$

## 7.5 Tests d'adéquation non paramétriques.

Nous allons maintenant nous tourner vers un problème de test plus complexe que celui de Neyman-Pearson, dans lequel en particulier nous ne chercherons pas à trouver un test optimal. Il nous faut d'abord introduire une nouvelle notion.

### 7.5.1 Niveaux asymptotiques, consistance.

**Définition 35** Soit  $\alpha$  fixé dans  $(0, 1)$ . On se donne une suite générale d'expériences de la forme,

$$\mathcal{E}_n = (\Omega_n, \mathcal{F}_n, X^n, \mathcal{X}_n, \mathcal{A}_n, P_\theta^n, \theta \in \Theta),$$

(échantillonnée ou non), et un problème de test associé à la partition  $\Theta_0, \Theta_1$ . On dira qu'une suite  $\Phi_n$  de tests, adaptée à  $\mathcal{E}_n$ , est **asymptotiquement de niveau  $\alpha$** , si

$$\forall \theta \in \Theta_0, \quad \mathbf{E}_\theta^n \Phi_n \longrightarrow \alpha.$$

Dans les problèmes complexes, faute d'avoir facilement des tests de niveau  $\alpha$ , on se contente, bien que ce soit effectivement moins précis, de tests, asymptotiquement de niveau  $\alpha$ .

**Définition 36** Dans le contexte ci-dessus, on dit que la suite de tests  $\phi_n$  asymptotiquement de niveau  $\alpha$ , est **consistante** si

$$\forall \theta \in \Theta_1, \quad \mathbf{E}_\theta^n \Phi_n \longrightarrow 1.$$

Nous allons maintenant présenter le problème de tester l'adéquation à une loi donnée : On dispose d'un  $n$ -échantillon et on veut tester  $P = P_0$  où  $P_0$  est une loi connue, mais contre  $P \neq P_0$ . Dans ce cas l'alternative est composée de toutes les probabilités sauf  $P_0$ , ce qui est souvent un ensemble qu'on ne peut pas inclure dans  $\mathbf{R}^k$ . Dans ce cas, il ne sera pas question de chercher d'optimalité (c'est en général impossible).

Nous allons proposer ici un test classique pour répondre à ce problème.

### 7.5.2 Test d'adéquation du $\chi^2$ .

Le cadre d'application du test du  $\chi^2$  est le suivant :

1. On observe un  $n$ -échantillon de variables aléatoires à valeurs dans un espace absolument arbitraire.
2.  $P_0$  est une loi arbitraire sur  $(\mathcal{X}, \mathcal{F})$ , connue, et on veut tester  $\mathcal{H}_0 : P = P_0$  contre  $P \neq P_0$ .
3. On opère un processus de mise en classe (qui peut d'ailleurs avoir été pré-effectué directement sur les données). C'est à dire que l'on considère une partition mesurable de  $\mathcal{X} : A_1, \dots, A_k$ , telle que  $P_0(A_j) > 0, \forall j$ .
4. On note la statistique de comptage de l'ensemble  $A_j$ ,

$$N_j = \sum_{i=1}^n 1_{A_j}(X_i)$$

On fabrique la statistique de test :

$$R_n = n \sum_{j=1}^k \left( \frac{N_j}{n} - P_0(A_j) \right)^2 \frac{1}{P_0(A_j)}$$

**Théorème 7** *Sous  $\mathcal{H}_0$ ,  $R_n$  converge en loi vers une variable  $Z$  qui suit une loi de  $\chi^2(k-1)$ .*

**Corollaire 4** *Soit  $\chi_\alpha(k-1)$  déterminé par :*

$$P(Z \geq \chi_\alpha(k-1)) = \alpha$$

*si  $Z$  suit une loi de  $\chi^2(k-1)$ . La suite de tests*

$$\Phi_n(x_1, \dots, x_n) = 1_{\{R_n \geq \chi_\alpha(k-1)\}}$$

*est asymptotiquement de niveau  $\alpha$  pour tester  $\mathcal{H}_0 : P = P_0$ .*

La preuve du corollaire est une conséquence immédiate de la convergence en loi.

### Preuve du Théorème.

Considérons le vecteur

$$Z_n = (Z_n^1, \dots, Z_n^k)^*,$$

$$\text{où } Z_n^j = \frac{1}{\sqrt{n}}(N_j - nP_0(A_j)).$$

On peut écrire  $Z_n$  sous la forme

$$Z_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n (Y_i - \mathbf{E}_0 Y_i)$$

Si

$$Y_i = (1_{A_1}(X_i), \dots, 1_{A_k}(X_i))^*.$$

Sous  $\mathcal{H}_0$ , les variables  $Y_i$  sont i.i.d., centrées, de matrice de covariance  $\Gamma$ .

$$\Gamma_{jl} = P_0(A_j)\delta_{jl} - P_0(A_j)P_0(A_l),$$

si  $\delta_{jl} = 1$  si  $j = l$ , 0 sinon.

On peut donc appliquer le théorème de la limite centrale/ Sous  $\mathcal{H}_0$ ,  $Z_n$  converge en loi vers une variable aléatoire  $Z$  de loi  $N_k(0, \Gamma)$ .

On remarque par ailleurs que

$$R_n = f(Z_n), \quad f(y_1, \dots, y_k) = \sum_{j=1}^k \frac{y_j^2}{P_0(A_j)}.$$

Comme  $f$  est une fonction continue, on déduit que sous  $\mathcal{H}_0$ ,  $f(Z_n)$  converge en loi vers une variable aléatoire  $f(Z)$ .

Il nous reste à étudier la loi de  $f(Z)$ . On remarque qu'on peut écrire

$$f(Z) = \|AZ\|^2$$

si  $A$  est la matrice, nulle en dehors de la diagonale et dont les coefficients diagonaux sont  $\frac{1}{\sqrt{P_0(A_j)}}$ . On a que  $AZ$  suit une loi  $N_k(0, P)$ , si  $P = A\Gamma A^*$ . On vérifie aisément que  $P$  est une matrice de projection orthogonale. (i.e.  $P$  est symétrique et  $P^2 = P$ ) De sorte que, en loi, sous  $\mathcal{H}_0$ , on peut écrire :

$$AZ = P\xi$$

où  $\xi$  est un vecteur gaussien standard de  $\mathbf{R}^k$ . On déduit alors de la proposition 6 que  $f(Z)$  suit une loi  $\chi^2(\text{rg}(P))$ .

Il reste à montrer que  $rg(P) = k - 1$  : Exercice (utiliser la décomposition  $y^*y = y^*Py + (\sum_{j=1}^k (y_j \sqrt{P_0(A_j)}))^2, \forall y \in \mathbf{R}^k$ ). ■

Nous allons étudier maintenant les propriétés de consistances de cette suite de tests :  
Introduisons l'alternative

$$\mathcal{H}'_1 = \{P, \text{ Probabilité sur } (\mathcal{X}, \mathcal{F}), \exists j, P(A_j) \neq P_0(A_j)\}$$

**Proposition 18** *La suite de test  $\phi_n$  introduite ci-dessus, est consistante pour tester  $\mathcal{H}_0$  contre  $\mathcal{H}'_1$ .*