Ecole Normale Supérieure
Gérard Biau
FIMFA — Statistique

CONTRÔLE TERMINAL — 5/01/2012

Durée 3h — Aucun document autorisé

Exercice 1

On considère la famille de densités $\{f_\theta : \theta \in \mathbb{R}\}$, par rapport à la mesure de Lebesgue sur $\mathbb{R}$, de la forme :

$$f_\theta(x) = \frac{1}{2} \exp\Big(- |x - \theta| \Big).$$

On observe n variables aléatoires $X_1, \ldots, X_n$ i.i.d. de densité commune f_θ et on désigne par X une variable aléatoire de densité f_θ.

1. Calculer $\mathbb{E}_\theta(X)$ et $\mathbb{V}_\theta(X)$.
2. Expliciter la fonction de répartition de X et en déduire la médiane de la loi de X.
3. Préciser $\hat{\theta}_n$, l'estimateur de θ par la méthode des moments.
4. Donner la loi limite de $\hat{\theta}_n$ lorsque $n \to +\infty$.
5. 5.a Calculer le risque quadratique de $\hat{\theta}_n$.
   5.b Soit $\alpha \in]0;1[$. Si n est trop petit pour que l'on puisse utiliser la loi limite, comment faites-vous pour construire un intervalle de confiance symétrique pour θ de niveau $1 - \alpha$? Donner précisément la forme de cet intervalle.

   On suppose désormais, et jusqu'à la question 7 incluse, que n est assez grand pour que l'on puisse utiliser la loi limite de l'estimateur comme approximation de sa vraie loi.
6. Construire un intervalle de confiance symétrique pour θ au niveau de confiance 96% (on rappelle que le quantile d'ordre $0,98$ d'une loi $\mathcal{N}(0,1)$ a pour valeur $z_{0,98} = 2,05$).

7. 7.a On veut tester l'hypothèse que $\theta \leq 0$ contre l'alternative $\theta > 0$ au niveau 2%. Donner la région de rejet du test.
   7.b Calculer la puissance $\beta_n(\theta)$ de ce test. Quelle valeur minimale de n faut-il choisir pour que la puissance soit supérieure ou égale à $0,98$ pour $\theta \geq 1/4$?
8. Un statisticien torturé souhaite estimer le paramètre $\sqrt{|\theta|}$. Proposer un estimateur de cette quantité et donner sa loi limite.

Exercice 2

On considère le modèle de régression

$$Y_i = \theta x_i^2 + \varepsilon_i, \quad i = 1, \ldots, n,$$

où les x_i sont **déterministes** et les erreurs ε_i sont indépendantes et distribuées selon une loi normale d'espérance 0 et de variance σ^2 connue.

1. Calculer $\hat{\theta}_n$, estimateur du maximum de vraisemblance de θ. Préciser son biais et sa variance.
2. Calculer l'information de Fisher $J(\theta)$ du modèle par rapport aux observations $Y_1, \ldots, Y_n$.
3. Soit $\alpha \in]0, 1[$. Donner un intervalle de confiance de niveau $1 - \alpha$ pour θ, basé sur la statistique $\hat{\theta}_n$.

Exercice 3

On veut garantir l'anonymat dans certaines enquêtes par sondage. Admettons que l'on cherche à savoir quelle est la proportion p de personnes qui remplissent leur déclaration fiscale de façon honnête. On demande alors à chaque personne interrogée de se retirer dans une pièce et de jouer à pile ou face de la façon suivante :

– **Si elle obtient "pile"**, elle doit répondre **honnêtement** par "oui" ou par "non" à la question : "Votre déclaration fiscale est-elle honnête ?"
– **Si elle obtient "face"**, elle doit **lancer la pièce à nouveau** et répondre par "oui" ou par "non" à la question : "Avez-vous obtenu "face" au second tirage ?"

Remarque : *si la personne interrogée répond "oui" (resp. "non"), il est impossible à l'enquêteur de savoir à quelle question la réponse est "oui" (resp. "non"). L'anonymat est donc préservé.*

On note alors p la proportion (théorique) de déclarations fiscales honnêtes dans la population considérée et π la proportion (théorique) de réponses "oui". On désigne enfin par X la variable aléatoire qui compte le nombre de "oui" dans une enquête menée sur n personnes.

1. Etablir une relation simple entre π et p.
2. Donner un estimateur $\hat{\pi}$ de π. En déduire un estimateur $\hat{p}$ de p.
3. Calculer le risque quadratique de $\hat{p}$.
4. Soit $\alpha \in]0,1[$. Donner un intervalle de confiance asymptotique de niveau $1-\alpha$ pour p.
5. **Application numérique** : $n = 1000$, $x = 600$. On rappelle que, pour une loi $\mathcal{N}(0,1)$, $z_{0,975} = 1,96$. Donner une estimation de p et un intervalle de confiance (asymptotique) à 95%. Quel est le prix payé pour la confidentialité ?

Exercice 4

1. On considère une fonction de répartition F_0 sur $\mathbb{R}$. On dispose de n observations i.i.d. de fonction de répartition commune inconnue F. Rappeler comment tester que $F = F_0$ contre l'alternative $F \neq F_0$ au niveau α.
2. On suppose dans cette question que F_0 est la fonction de répartition associée à la densité f_0 de l'exercice 1.
   2.a Si G est la fonction de répartition de la loi uniforme sur $[-1;1]$, que vaut
   $$\Delta = \sup_x |F_0(x) - G(x)| \quad ?$$
   2.b En déduire une minoration de la puissance du test précédent lorsque $F = G$ et n est suffisamment grand.

Problème

Dans tout le devoir, la lettre $\mathcal{B}$ désigne la tribu borélienne de $\mathbb{R}^d$. On rappelle que si f et g sont deux densités de probabilité sur $\mathbb{R}^d$, c'est-à-dire deux fonctions mesurables positives ou nulles telles que
$$\int f = \int g = 1,$$

(toutes les intégrales sont calculées par rapport à la mesure de Lebesgue), alors

$$\int |f-g| = 2\sup_{B\in\mathcal{B}}\left|\int_B f - \int_B g\right| = 2\int_{A_{fg}}(f-g),$$

où A_{fg} est l'ensemble $\{f > g\}$, c'est-à-dire

$$A_{fg} = \left\{x \in \mathbb{R}^d : f(x) > g(x)\right\}.$$

Ce résultat est connu sous le nom de *théorème de Scheffé*.

A. Préliminaires. Soit $\mathcal{A}$ une classe de sous-ensembles de $\mathbb{R}^d$, de cardinal (pas nécessairement fini) > 1. On appelle *coefficient de pulvérisation* de n points par la classe $\mathcal{A}$ la quantité

$$\mathbf{S}_{\mathcal{A}}(n) = \max_{x_1,\ldots,x_n\in\mathbb{R}^d} \text{Card}\left\{\{x_1,\ldots,x_n\}\cap A : A\in\mathcal{A}\right\}.$$

La *dimension de Vapnik-Chervonenkis* $V_{\mathcal{A}}$ de $\mathcal{A}$ est alors définie comme le plus grand entier n tel que

$$\mathbf{S}_{\mathcal{A}}(n) = 2^n$$

(lorsque $\mathbf{S}_{\mathcal{A}}(n) = 2^n$ pour tout n, on pose $V_{\mathcal{A}} = +\infty$). Lorsque $V_{\mathcal{A}} < +\infty$, on admettra le résultat combinatoire suivant, connu sous le nom de *lemme de Sauer* :

$$\mathbf{S}_{\mathcal{A}}(n) \leq \sum_{i=0}^{V_{\mathcal{A}}} \binom{n}{i}.$$

1. Que représentent $\mathbf{S}_{\mathcal{A}}(n)$ et $V_{\mathcal{A}}$?
2. **Exemple**. Si $d = 1$ et $\mathcal{A} = \{]-\infty; a] : a \in \mathbb{R}\}$, que vaut $V_{\mathcal{A}}$? Et si $\mathcal{A} = \{[a; b] : a, b \in \mathbb{R}\}$?
3. Soit $\mathcal{A}$ une classe d'ensembles de cardinal fini > 1. Montrer que

$$V_{\mathcal{A}} \leq \log_2(\text{Card } \mathcal{A}).$$

4. Soit $\mathcal{A}$ une classe d'ensembles admettant une dimension de Vapnik-Chervonenkis $V_{\mathcal{A}}$ finie. Prouver que, pour tout n,

$$\mathbf{S}_{\mathcal{A}}(n) \leq (n+1)^{V_{\mathcal{A}}}.$$

5. Soit $X_1, \ldots, X_n$ un échantillon d'observations i.i.d. dont la loi commune est μ, et soit $\mathcal{A}$ une classe de boréliens de $\mathbb{R}^d$ de cardinal **fini**. Montrer alors que

$$\mathbb{E}\left\{\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)|\right\} \leq \sqrt{\frac{\log(2e \times \text{Card } \mathcal{A})}{2n}},$$

où μ_n représente la mesure empirique associée à $X_1, \ldots, X_n$ (c'est-à-dire $\mu_n(B) = (1/n)\sum_{i=1}^n \mathbf{1}_{[X_i\in B]}$).

Dans toute la suite du problème, nous admettrons que si $\mathcal{A}$ est une classe **quelconque** de boréliens de $\mathbb{R}^d$, alors

$$\mathbb{E}\left\{\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)|\right\} = \mathcal{O}\left(\sqrt{\frac{\log\left(\mathbf{S}_{\mathcal{A}}(n)\right)}{n}}\right).$$

D'une certaine façon, ce résultat généralise donc celui de la question 5.

B. Un problème de sélection. On se donne désormais un échantillon $X_1, \ldots, X_n$ de variables aléatoires indépendantes et de même loi à densité **inconnue** f. On se donne également un ensemble $\mathcal{F}$ de densités candidates, paramétrées par un paramètre θ :

$$\mathcal{F} = \{f_\theta : \theta \in \Theta\}.$$

Le but du jeu consiste à sélectionner dans $\mathcal{F}$ la "meilleure" densité possible, en se basant exclusivement sur l'échantillon $X_1, \ldots, X_n$.

1. Soit μ_n la mesure empirique associée à l'échantillon $X_1, \ldots, X_n$. Expliquer pourquoi la stratégie consistant à sélectionner θ dans Θ en minimisant en θ la quantité

$$\sup_{B\in\mathcal{B}} \left|\int_B f_\theta - \mu_n(B)\right|$$

est une fausse bonne idée.

2. On introduit alors la collection d'ensembles

$$\mathcal{A} = \left\{\{f_\theta > f_{\theta'}\} : (\theta, \theta') \in \Theta^2\right\}.$$

Afin de choisir la "meilleure" densité dans $\mathcal{F}$, on propose désormais de minimiser en θ le critère suivant :

$$\Delta(\theta) = \sup_{A\in\mathcal{A}} \left|\int_A f_\theta - \mu_n(A)\right|.$$

On note θ^* un élément de Θ tel que $\Delta(\theta^*) = \inf_{\theta\in\Theta} \Delta(\theta)$ (on suppose pour simplifier que θ^* existe).

2.a Soit $\bar{\theta}$ un élément de Θ tel que

$$\int |f_{\bar{\theta}} - f| = \inf_{\theta\in\Theta} \int |f_\theta - f|$$

(on suppose encore que $\bar{\theta}$ existe). Montrer que

$$\int |f_{\theta^*} - f_{\bar{\theta}}| \leq 4 \sup_{A\in\mathcal{A}} \left| \int_A f_{\bar{\theta}} - \mu_n(A) \right|.$$

2.b Montrer alors que

$$\int |f_{\theta^*} - f| \leq 3 \inf_{\theta\in\Theta} \int |f_\theta - f| + 4\Delta_n,$$

où Δ_n est une quantité aléatoire que l'on précisera.

3. 3.a On désigne par $\mathbf{S}_{\mathcal{A}}(n)$ le coefficient de pulvérisation de n points par la classe $\mathcal{A}$. Montrer que

$$\mathbb{E}\left\{ \int |f_{\theta^*} - f| \right\} \leq 3 \inf_{\theta\in\Theta} \int |f_\theta - f| + \mathcal{O}\left(\sqrt{\frac{\log\left(\mathbf{S}_{\mathcal{A}}(n)\right)}{n}} \right).$$

3.b Donner une interprétation statistique de l'inégalité ci-dessus.

C. Application. On se place sur $\mathbb{R}$ et on choisit comme classe $\mathcal{F}$ l'ensemble des densités gaussiennes paramétrées par leur moyenne et leur variance, c'est-à-dire

$$\mathcal{F} = \left\{ f_\theta(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-(x-m)^2/(2\sigma^2)} : \theta = (m, \sigma^2) \in \mathbb{R} \times]0; +\infty[\right\}.$$

1. Montrer que, dans ce cas, $\mathcal{A}$ est contenue dans une classe d'ensembles $\mathcal{B}_2$ que l'on décrira.
2. Déterminer la dimension de Vapnik-Chervonenkis V de $\mathcal{B}_2$.
3. En déduire que

$$\mathbb{E}\left\{ \int |f_{\theta^*} - f| \right\} \leq 3 \inf_{\theta\in\Theta} \int |f_\theta - f| + \mathcal{O}\left(\sqrt{\frac{V \log n}{n}} \right).$$