ENSAE deuxième année
Examen, session 2003, 2 heures

SE 203 : ESTIMATION ET INTRODUCTION AUX TESTS

Paul Doukhan

Les documents distribués en cours et les notes manuscrites. Les calculettes sont inutiles et interdites. On pourra aussi se référer librement aux lemmes, théorèmes et propositions donnés dans le polycopié support de cours à condition d'y référer avec précision.
On traitera les exercices dans un ordre arbitraire, à la condition de le préciser clairement.

La longueur intentionnellement excessive de l'énoncé est compensée par un barême portant sur plus de 20 points.
Barême indicatif sur 26 points *: A(1,1,1,1,1,2,2) sur 9 points, B(1,1,1,2) sur 5 points, C(1,1,1,1,1) sur 5 points, D(3,2,2) sur 7 points.*

Exercice A *(lois* Γ*)* Soit $X = (X_1, \ldots, X_n)$ un n-échantillon de loi $\Gamma(p, \lambda)$ de densité $g_{p,\lambda}(x) = \lambda^p x^{p-1} e^{-\lambda x}/\Gamma(p)$ sur $\mathbb{R}^+$. On rappelle ici que la fonction d'Euler, $\Gamma(p) = \int_0^\infty t^{p-1}e^{-t}dt$ est régulière sur $]0, \infty[$ et vérifie $\Gamma(p+1) = p\Gamma(p)$ pour $p > 0$ et $\Gamma(1) = 1$.

1. Montrez que X est distribué selon un modèle exponentiel canonique en dimension 2. Précisément, on montrera que sa densité (par rapport à la la mesure de Lebesgue sur $(\mathbb{R}^+)^n$) s'écrit $h(x)e^{T(x)\cdot\theta - A(\theta)}$ avec $\theta = (\theta_1, \theta_2)$ pour $\theta_1 = p$, $\theta_2 = -\lambda$ et on explicitera $h(x)$, $T(x)$, et $A(\theta)$.
2. Calculez l'information de Fisher de ce modèle (on pourra donner une expression dans laquelle figurent des dérivées de la fonction Γ d'Euler).
3. Prouvez que les équations du maximum de vraisemblance équivalent à

$$\frac{\Gamma'}{\Gamma}(\widehat{p}) - \log\widehat{\lambda} = \overline{\log X}, \qquad \frac{\widehat{p}}{\widehat{\lambda}} = \overline{X}$$

   Devant la haute non linéarité de cette expression, on envisage des méthodes de résolution numériques. Pour initialiser ces méthodes, nous mettons en oeuvre les méthodes de moments dans la suite de cet exercice.
4. Donnez une expression de $\mu_k = \mathbb{E}_\theta X_1^k$. On pourra expliciter cette expression lorsque $k = 1, 2, 3$
5. Déterminez des estimateurs de θ fondés sur les moments d'ordre 1 et 2.
6. Déterminez des estimateurs de θ fondés sur les moments d'ordre 2 et 3.
7. Que peut-on dire des propriétés asymptotiques de ces estimateurs ? (le barême de cette question ouverte pourra varier selon la précision de la réponse.)

Exercice B *(information d'un modèle de Poisson)* Soit $n \geq 1$ entier, $z_1, \ldots, z_n \in \mathbb{R}$, on observe des variables aléatoires indépendantes $X_1, \ldots, X_n$ telles que X_i suive la loi de Poisson $\mathcal{P}(\mu_i)$ de paramètre $\mu_i = e^{\alpha+\beta z_i}$. Par exemple z_i pourrait modéliser la quantité de médicament donnée au patient i et Y_i mesurerait le nombre d'agents infectieux par unité de sang de ce patient 24 heures après l'absorption du médicament. Ici $\alpha, \beta \in \mathbb{R}$ sont des paramètres inconnus.

1. Montrez que $X = (X_1, \ldots, X_n)$ est distribué selon un modèle exponentiel canonique de paramètre $\theta = (\alpha, \beta)$ et de rang 2 que l'on précisera.
2. Calculez la matrice d'information de ce modèle.
3. Déterminez des bornes inférieures de la variance d'estimateurs (réguliers) sans biais de α et β.
4. Lorsque $z_i = i/(n+1)$ pour $1 \leq i \leq n$ donnez des équivalents de ces bornes en utilisant l'approximation d'une intégrale à l'aide de sommes de Riemann (la notation dépendra ici de la rigueur de la réponse proposée).

Exercice C *(intervalles de confiance et estimateurs)* Soit $\xi_1, \ldots, \xi_n$ une suite indépendante et de loi $\mathcal{N}(0, \sigma^2)$ pour un $\sigma^2 > 0$ connu. On suppose que l'on observe $X_i = \frac{\theta}{2} z_i^2 + \xi_i$ pour un paramètre $\theta \in \mathbb{R}$ inconnu et où le plan d'expérience $z_1, \ldots, z_n \in \mathbb{R}$ est fixé.

1. Soit $\widehat{\theta}$ un estimateur gaussien sans biais de θ dont la variance v est indépendante de θ. Soit $\alpha > 0$, fixé, déterminez un intervalle de confiance pour θ au niveau $1 - \alpha$. Comment choisir cet intervalle pour que sa longueur soit minimale ?
2. Calculez l'estimateur du maximum de vraisemblance $\widehat{\theta}_1$ de θ. Montrez qu'il est sans biais et calculez sa variance. On prouvera que cette variance v_1 est indépendante de θ.
3. Montrez que $\widehat{\theta}_2 = 2\frac{\sum_{i=1}^n X_i}{\sum_{i=1}^n z_i^2}$ estime θ sans biais. Prouvez que sa variance v_2 est indépendante de θ.
4. Proposez un intervalle de confiance raisonnable $I_n = [a_n, b_n]$ de θ au niveau $1 - \alpha$; on pourra commencer par comparer v_1 et v_2.
5. Donnez un équivalent asymptotique de $b_n - a_n$ lorsque $z_i = i/(n+1)$ pour $1 \leq i \leq n$.

Exercice D *(test gaussien)* Soit $X = (X_1, \ldots, X_n)$ un n-échantillon de loi normale $\mathcal{N}(0, 1/\theta)$.

1. Calculer $\hat{\theta}$ l'estimateur de maximum de vraisemblance de θ. Construire un estimateur sans biais de θ. Cet estimateur est-il efficace, asymptotiquement efficace ?
2. Construire le test le plus puissant de $\theta = 1$ contre $\theta > 1$. Si $n = 15$ et si $\sum_{i=1}^{15} x_i^2 = 6,8$, effectuer le test au niveau 5%. Peut-on avoir une idée de la fonction puissance de ce test ?
3. Construire le test le plus puissant de $\theta = 1$ contre $\theta \neq 1$. Effectuer le test au niveau 5% pour les mêmes données qu'au point précédent.

SE 203 : Corrigé succinct :
estimation et introduction aux tests

Paul Doukhan

Exercice A *(lois Γ)* Soit $X = (X_1, \ldots, X_n)$ un n-échantillon de loi $\Gamma(\lambda, p)$ de densité $g_{p,\lambda}(x) = \lambda^p x^{p-1} e^{-\lambda x}/\Gamma(p)$ sur $\mathbb{R}^+$, où on rappelle que $\Gamma(p) = \int_0^\infty t^{p-1} e^{-t} dt$ vérifie $\Gamma(p+1) = p\Gamma(p)$ pour $p > 0$ et $\Gamma(1) = 1$.

1. Ici
$$T = \left(\sum_{i=1}^n \log x_i, \sum_{i=1}^n x_i \right), \; h(x) = \frac{1}{x_1 \cdots x_n}, \; A(\theta_1, \theta_2) = n(\log \Gamma(\theta_1) - \theta_1 \log(-\theta_2)).$$

2. $I_n(\theta) = nI_1(\theta)$ où
$$I_1(\theta) = \begin{pmatrix} \frac{\Gamma''}{\Gamma}(\theta_1) - \frac{\Gamma'^2}{\Gamma^2}(\theta_1) & \frac{1}{\theta_2} \\ \frac{1}{\theta_2} & -\frac{\theta_1}{\theta_2^2} \end{pmatrix}.$$

3. Revenant à la défnition, on prouve que
$$\frac{\Gamma'}{\Gamma}(\widehat{p}) - \log \widehat{\lambda} = \overline{\log X}, \qquad \frac{\widehat{p}}{\widehat{\lambda}} = \overline{X}$$

4. $$\mu_k = \int_0^\infty x^k g_{p,\lambda}(x) dx = \frac{\Gamma(p+k)}{\lambda^k \Gamma(p)} \int_0^\infty g_{p+k,\lambda}(x) dx = \frac{\Gamma(p+k)}{\lambda^k \Gamma(p)}$$
$$\mu_1 = \frac{p}{\lambda}, \; \mu_2 = \frac{p(p+1)}{\lambda^2}, \; \mu_3 = \frac{p(p+1)(p+2)}{\lambda^3}$$

5. Ici posant $\sigma^2 = \mu_2 - \mu_1^2$, il vient $\mu_1 = \frac{p}{\lambda}$ et $\sigma^2 = \frac{p}{\lambda^2}$ donc $\lambda = \mu_1/\sigma^2$ et $p = \mu_1^2/\sigma^2$ sont estimés empiriquement en injectant dans les relations précédentes $\widehat{\mu_1} = \frac{1}{n}\sum_{i=1}^n X_i$ et $\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^n X_i^2 - \left(\widehat{\mu_1}\right)^2$.

6. On note qu'ici $\mu_2(p+2) = \mu_3\lambda$ pour en déduire que $p(p+1)\mu_3^2 = (p+2)^2\mu_2^3$ ainsi p est solution de l'équation du second degré
$$(\mu_3^2 - \mu_2^3)p^2 + (\mu_3^2 - 4\mu_2^3)p - 4\mu_2^3 = 0$$
cette équation a pour discriminant $\Delta = \mu_3^4 + 8\mu_2^3\mu_3^2 \geq 0$ et on vérifie que $p = \frac{\mu_2^3 - \mu_3^2 + \sqrt{\mu_3^4 + 8\mu_2^3\mu_3^2}}{2\left(\mu_3^2 - \mu_2^3\right)}$.

7. Lois des grands nombres et TLC s'appliquent via la Delta méthode.

Exercice B *(information d'un modèle de Poisson)*

1. Ici, $T(x) = (\sum_{i=1}^n x_i, \sum_{i=1}^n z_i x_i)$ et $A(\theta) = \sum_{i=1}^n \mu_i = \sum_{i=1}^n e^{\alpha+\beta z_i}$.
2. $$I = D^2 A(\theta) = \begin{pmatrix} \sum_{i=1}^n e^{\alpha+\beta z_i} & \sum_{i=1}^n z_i e^{\alpha+\beta z_i} \\ \sum_{i=1}^n z_i e^{\alpha+\beta z_i} & \sum_{i=1}^n z_i^2 e^{\alpha+\beta z_i} \end{pmatrix}$$
3. $$\operatorname{Var} \widehat{\alpha} \geq \left(\sum_{i=1}^n e^{\alpha+\beta z_i} \right)^{-1}, \operatorname{Var} \widehat{\beta} \geq \left(\sum_{i=1}^n z_i^2 e^{\alpha+\beta z_i} \right)^{-1}$$
4. Pour tout $p \geq 0$, on a

$$\begin{aligned} t_p &= \frac{1}{n} \sum_{i=1}^n z_i^p e^{\alpha+\beta z_i} \\ &= \frac{e^\alpha}{n} \sum_{i=1}^n \log^p \left(\frac{i}{n+1} \right) \left(\frac{i}{n+1} \right)^\beta \\ &\sim e^\alpha \int_0^1 \log^p x x^\beta dx \end{aligned}$$

pour la dernière équivalence notons que la fonction $x \mapsto \log^p x x^\beta$ est monotone et intégrable au voisinage de 0 ce qui permet d'affirmer la convergence des sommes de Riemann précédentes lorsque $p \geq 0$.

Lemme 1 *Soit* $f :]0,1] \to \mathbb{R}^+$ *une fonction décroissante, de restriction à tout intervalle* $[a, 1]$ *(*$0 < a < 1$*) Riemann intégrable. Alors si l'intégrale généralisée* $\int_0^1 f(t)dt$ *converge au sens que* $\lim_{a \downarrow 0} \int_0^1 f(t)dt$ *existe, on a*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^n f\left(\frac{i}{n}\right) = \int_0^1 f(t)dt$$

Preuve du lemme (non demandée). Par monotonie

$$\int_{(i)/n}^{(i+1)/n} f(t)dt \leq \frac{1}{n} f\left(\frac{i}{n}\right) \leq \int_{(i-1)/n}^{i/n} f(t)dt$$

donc

$$\int_0^{1-1/n} f(t)dt - \frac{1}{n} f\left(\frac{1}{n}\right) \leq \frac{1}{n} \sum_{i=1}^n f\left(\frac{i}{n}\right) \leq \int_0^1 f(t)dt$$

Pour conclure on note que

$$\frac{1}{n} f\left(\frac{1}{n}\right) = 2\frac{1}{2n} f\left(\frac{1}{n}\right) \leq \int_{1/(2n)}^{1/n} f(t)dt \to_{n\to\infty} 0.$$

L'intégrale $\int_0^1 \log^p x x^\beta dx$ est calculée par récurrence grâce à des intégrations par parties. Lorsque $p=0$ on obtient, par exemple $t_0 \sim e^\alpha/(\beta+1)$.

Le calcul précédent prouve que les bornes recherchées sont $(nt_0)^{-1}$ et $(nt_2)^{-1}$.

Exercice C *(intervalles de confiance et estimateurs)* Soit $\xi_1, \ldots, \xi_n$ une suite indépendante et de loi $\mathcal{N}(0,\sigma^2)$ pour un $\sigma^2 > 0$ connu. On suppose que l'on observe $X_i = \frac{\theta}{2} z_i^2 + \xi_i$ pour un paramètre $\theta \in \mathbb{R}$ inconnu et et où le plan d'expérience $z_1, \ldots, z_n \in \mathbb{R}$ est fixé.

1. Soit $\widehat{\theta}$ un estimateur gaussien sans biais de θ dont la variance v est indépendante de θ. Soit $\alpha > 0$, fixé, déterminez un intervalle de confiance pour θ au niveau $1-\alpha$. Comment choisir cet intervalle pour que sa longueur soit minimale. Soient $a < b$ tels que $\mathbb{P}(N \in [a,b]) = 1-\alpha$ pour $N \sim \mathcal{N}(0,1)$, alors $\mathbb{P}\left(\frac{\widehat{\theta}-\theta}{\sqrt{v}} \in [a,b]\right) = 1-\alpha$.
   Par suite $[\widehat{\theta} - \sqrt{v}b, \widehat{\theta} - \sqrt{v}a]$ est un intervalle de confiance pour θ au niveau $1-\alpha$. On choisit alors $a = -b = \varphi_{1-\alpha/2}$, des quantiles de la gaussienne standard.
2. On calcule
$$\widehat{\theta}_1 = 2\frac{\sum_{i=1}^n z_i^2 X_i}{\sum_{i=1}^n z_i^4}$$
   et
$$v_1 = \frac{4\sigma^2}{\sum_{i=1}^n z_i^4}$$
3. $$v_2 = \frac{4n\sigma^2}{\left(\sum_{i=1}^n z_i^2\right)^2}$$
4. Par l'inégalité de Cauchy-Schwartz, $v_1 \le v_2$. Par suite on posera
$$I = [\widehat{\theta}_1 + \sqrt{v_1}\varphi_{1-\alpha/2}, \widehat{\theta} - \sqrt{v_1}\varphi_{1-\alpha/2}]$$
5. On utilise encore des sommes de Riemann.

Exercice D *(test gaussien)*

1. $\hat{\theta} = n/\sum_{i=1}^n X_i^2$. Notons que $T = \sum_{i=1}^n X_i^2 \sim \frac{1}{\theta}\Gamma(n/2, 1/2)$ donc $\mathbb{E}_\theta n T^{-1} = n\theta \mathbb{E} 1/\gamma$ si $\gamma \sim \Gamma(n/2,1/2)$ et avec le calcul de l'exercice A, $\mathbb{E}_\theta n T^{-1} = n\theta\mathbb{E}1/\gamma = n\theta\mathbb{E}1/(n-2)$. Alors $\tilde{\theta} = \frac{n-2}{n}\hat{\theta} = \frac{n-2}{T}$ estime θ sans biais. Il ne peut être efficace car $\frac{1}{\theta}$ n'est pas fonction affine de θ. On prouve, par contre qu'il est asymptotiquement efficace.
2. Soit $\theta_1 < \theta_2$ la fonction $V(x) = p_{\theta_2}(x)/p_{\theta_1}(x) = (\theta_2/\theta_1)^{n/2} e^{-T(\theta_2-\theta_1)/2}$ est décroissante, donc la transposition du théorème 7.1 du cours à ce cas donne un test le plus puissant de $\theta = 1$ contre $\theta > 1$ de zone de rejet $(T < c)$.
   Si $n = 15$ et si $\sum_{i=1}^{15} x_i^2 = 6,8$, on effectue le test en lisant la table à l'envers $\mathbb{P}(\gamma > 7,21) = 0,95$ donne bien $\mathbb{P}(\gamma < 7,21) = 0,05$. Le test accepte donc l'expérience au niveau 5% mais pas à 2,5%.
3. La zone de rejet est ici $T < c_1$ ou $T > c_2$ pour $c_1 < c_2$ et $\mathbb{P}(\gamma < c_1) + \mathbb{P}(\gamma > c_2) = 0,05$. On choisit par exemple ces deux quantités égales pour mettre le test en oeuvre.