École Normale Supérieure
FIMFA — Statistique 2016–2017

Partiel

1. Pour θ in $\mathbb{R}$, la loi P_θ est définie par sa densité $p_\theta(x) = \frac{1}{2}\exp(-|x-\theta|)$.
   (a) (1 point) Calculer l'espérance et la variance de P_θ.

Solution: *Pour répondre à cette question, il faut essayer de réduire les calculs au minimum. La première observation utile : il n'est pas nécessaire de calculer espérance et variance en tout θ. Il suffit d'étuier le cas $\theta = 0$.*

Nous avons affaire à un modèle de localisation pur : si $X \sim P_0$ alors $X + \theta \sim P_\theta$. L'espérance sous P_θ est égale à θ plus l'espérance sous P_0. La variance ne dépend pas de θ.

Tous les moments polynomiaux et certains moments exponentiels sont définis pour tout θ.

Comme la densité de P_0 est une fonction paire, si $X \sim P_0$, X est *symétrique*, de même loi que $-X$, ce qui entraîne $\mathbb{E}_0 X = 0$ et

$$\mathbb{E}_\theta X = \theta.$$

La variance ne dépend pas du paramètre de localisation.

$$\operatorname{var}_\theta(X) = \operatorname{var}_0(X) = \int_0^\infty x^2 \mathrm{e}^{-x} \mathrm{d}x = 2\,.$$

   (b) (1 point) Proposer un estimateur de θ fondé sur la méthode des moments. Etudier sa consistance, sa normalité asymptotique.

Solution: On propose $\widehat{\theta}_n = \overline{X}_n$ (la moyenne empirique). D'après la réponse à la question précédente $\mathbb{E}_\theta \overline{X}_n = \theta$, il s'agit d'un estimateur sans biais dont le risque quadratique coïncide avec la variance $\operatorname{var}_\theta(\widehat{\theta}_n) = \frac{2}{n}$.

La suite $(\widehat{\theta}_n)_n$ est (fortement) consistante (loi des grands nombres). Et le TCL se traduit immédiatement en

$$\sqrt{n}(\widehat{\theta}_n - \theta) \rightsquigarrow \mathcal{N}(0,2)\ .$$

   (c) (1 point) La fonction score notée $\dot{\ell}_n$ est définie comme la dérivée par rapport à θ de la log vraisemblance quand cette dérivée est bien définie, 0 dans les autres cas. Si elles sont bien définies, que valent $\mathbb{E}_\theta[\dot{\ell}_n(\theta)]$ et $\mathbb{E}_\theta[\dot{\ell}_n(\theta)^2]$?

Solution: Avec probabilité 1, les points d'un n-échantillon d'une loi P_θ sont deux à deux distincts. Si θ est distinct de tous les X_i,

$$\dot{\ell}_n(\theta) = \sum_{i=1}^n \mathbb{I}_{X_i > \theta} - \mathbb{I}_{X_i < \theta}$$

Sous P_θ, $\dot{\ell}_n(\theta)$ est distribuée comme une somme de variables aléatoires indépendantes symétriques centrées à valeurs dans $\{-1, 1\}$. D'où

$$\mathbb{E}_\theta\left[\dot{\ell}_n(\theta)\right] = 0$$

et

$$\operatorname{var}_\theta\left[\dot{\ell}_n(\theta)\right] = n\,.$$

   (d) (2 points) L'estimateur au maximum de vraisemblance est-il bien défini dans ce modèle ? Si oui, étudier sa consistance et éventuellement sa normalité asymptotique.

Solution: L'estimateur du maximum de vraisemblance est une médiane de l'échantillon. Si l'échantillon est de taille impaire, il est défini de manière unique et vaut $X_{((n+1)/2,n)}$. Si l'échantillon est de taille paire, n'importe quelle valeur comprise entre $X_{(n/2,n)}$ et $X_{(n/2+1,n)}$ convient.

Si $\widetilde{\theta}_n$ désigne un estimateur au maximum de vraisemblance, on a d'après les propriétés de la médiane empirique vues en DM,

$$\text{sous } P_\theta \qquad \widetilde{\theta}_n \xrightarrow{P} \theta \qquad \text{et} \qquad \sqrt{n}\left(\widetilde{\theta}_n - \theta\right) \rightsquigarrow \mathcal{N}(0,1)\,.$$

L'estimateur au maximum de vraisemblance est sans biais (pour les tailles d'échantillon impaires), consistant, asymptotiquement normal, de variance asymptotique 1.

(e) (1 point) On veut tester $H_0 : \theta \leq 0$ contre $H_1 : \theta > \delta > 0$. Proposer un test de niveau asymptotique $\alpha > 0$ pour séparer ces deux hypothèses.

Solution: Comparer la médiane empirique à un seuil, rejeter H_0 si la médiane empirique est supérieure à ce seuil, en pas rejeter sinon, c'est équivalent à réaliser un test de rapport de vraisemblance généralisé où on compare à un seuil la différence entre la log-vraisemblance sous P_δ et la log-vraisemblance sous P_0.

Si $\theta < \theta'$, la fonction de répartition de la loi de la médiane empirique sous P_θ est supérieure ou égale à la fonction de répartition de la loi de la médiane empirique sous $P_{\theta'}$ (très facile à vérifier à l'aide d'un argument de couplage quantile). Il suffit pou calibrer le niveau asymptotique du test de s'intéresser à ce qui se passe en $\theta = 0$.

Si on choisit comme région critique l'événement

$$\{\widehat{\theta}_n \geq \sqrt{n} z_\alpha\}$$

avec ζ_α quantile d'ordre $1-\alpha$ de $\mathcal{N}(\prime, \infty)$, on obtient un test de niveau asymptotique α.

2. Dans un modèle exponentiel minimal, on considère deux paramètres θ et $\theta_n = \theta + h/\sqrt{n}$ où h verifie $\sum_{i=1}^d h_i = 0$.

(a) (1 point) Calculer la distance de Hellinger entre P_θ et P_{θ_n}. Donner un équivalent de $H^2(P_\theta, P_{\theta_n})$ lorsque $n \to \infty$.

Solution: On suppose sans perdre en généralité que le modèle est dominé par P_0. La densité de P_θ en x s'écrit alors $\exp\left(\langle \theta, T(x)\rangle - A(\theta)\right)$. L'affinité de Hellinger entre P_θ et $P_{\theta'}$ s'écrit

$$\rho(\theta,\theta') = \exp\left(A\left(\frac{\theta+\theta'}{2}\right) - \frac{A(\theta)+A(\theta')}{2}\right)\,.$$

On rappelle que $\nabla^2 A(\theta) = I(\theta)$. Lorsque $n \to \infty$, $1 - \rho(\theta_n, \theta)$ est équivalent à

$$\frac{A(\theta)+A(\theta_n)}{2} - A\left(\frac{\theta+\theta_n}{2}\right) \sim \frac{1}{8n} h^t I(\theta) h\,.$$

(b) (2 points) Quelle est la loi limite de $\ell_n(\theta_n) - \ell_n(\theta)$ sous P_θ, sous P_{θ_n} ?

Solution: Nous allons mettre en évidence le phénomène de normalité asymtotique locale (LAN property). Ce n'est pas une spécificité des modèles exponentiels. On observe ce phénomène dans les modèles réguliers (différentiables en moyenne quadratique).

$$\begin{aligned}
\ell_n(\theta_n) - \ell_n(\theta) &= n\langle \theta_n - \theta, \overline{T}_n\rangle + n\left(A(\theta) - A(\theta_n)\right) \\
&= \sqrt{n}\langle h, \overline{T}_n\rangle + n\left(A(\theta) - A(\theta + \frac{h}{\sqrt{n}})\right) \\
&= \langle h, \sqrt{n}\left(\overline{T}_n - \mathbb{E}\overline{T}_n\right)\rangle + \sqrt{n}\langle h, \mathbb{E}\overline{T}_n\rangle + n\left(A(\theta) - A(\theta + \frac{h}{\sqrt{n}})\right)
\end{aligned}$$

Sous P_θ, $\mathbb{E}\overline{T}_n = \nabla A(\theta)$, et $\sqrt{n}(\overline{T}_n - \mathbb{E}\overline{T}_n) \rightsquigarrow \mathcal{N}(0, \nabla^2 A(\theta)) = \mathcal{N}(0, I(\theta))$,

$$
\begin{aligned}
& +\sqrt{n}\langle h, \mathbb{E}\overline{T}_n\rangle + n\left(A(\theta) - A(\theta + \frac{h}{\sqrt{n}})\right) \\
= & -n\left(A(\theta + \frac{h}{\sqrt{n}}) - A(\theta) - \langle \frac{h}{\sqrt{n}}, \nabla A(\theta)\right) \\
= & -\frac{1}{2}h^T \nabla^2 A(\theta) h - \|h\|^2 R\left(\frac{h}{\sqrt{n}}\right)
\end{aligned}
$$

où $\lim_{x\to 0} R(x) = 0$. On peut conclure que cette expression converge vers $-\frac{1}{2}h^T\nabla^2 A(\theta)h = -\frac{1}{2}h^T I(\theta)h$ lorsque n tend vers l'infini.
Sous P_θ,

$$\ell_n(\theta_n) - \ell_n(\theta) \rightsquigarrow \mathcal{N}\left(-\frac{1}{2}h^T I(\theta)h, h^T I(\theta)h\right).$$

Le fait que l'espérance de la gaussienne limite soit l'opposée de la moitié de la variance de cette gaussienne limite se comprend intuitivement :

$$\mathbb{E}_\theta\left[\exp\left(\ell_n(\theta_n) - \ell_n(\theta)\right)\right] = 1.$$

On s'attend à ce que l'exponentielle de la gaussienne limite soit d'espérance 1. Ceci implique très précisément la relation entre variance et espérance.
Sous P_{θ_n}, en invoquant le TCL (version Lindeberg-Feller),

$$\ell_n(\theta_n) - \ell_n(\theta) \rightsquigarrow \mathcal{N}\left(\frac{1}{2}h^T I(\theta)h, h^T I(\theta)h\right).$$

Intuitivement, la relation entre espérance et variance de la loi limite se comprend comme auparavant.

(c) (2 points) Quelle est la puissance optimale limite de tests de niveau α entre P_θ et P_{θ_n} ?

Solution: On peut raisonner sur les tests de rapport de vraisemblance. Pour obtenir un niveau asymptotique α, il faut comparer $\ell_n(\theta_n) - \ell_n(\theta)$ au quantile d'ordre $1-\alpha$ de $\mathcal{N}\left(-\frac{1}{2}h^T I(\theta)h, h^T I(\theta)h\right)$ soit

$$-\frac{1}{2}h^T I(\theta)h + z_\alpha\sqrt{h^T I(\theta)h},$$

Si ce seuil est dépassé, on rejette l'hypothèse nulle θ.
La puissance limite de ce test sous l'alternative est

$$\overline{\Phi}\left(z_\alpha - \sqrt{h^T I(\theta)h}\right).$$

(d) (1 point) Sous P_θ, quelle est la limite en loi de $I(\theta)^{-1}\frac{1}{\sqrt{n}}\nabla\ell_n(\theta) + \sqrt{n}(\widehat{\theta}_n - \theta)$?

Solution: 0

3. Modèles de mélange.
On observe n tirages indépendants $X_1, \ldots, X_n$ selon une loi inconnue P sur $\{0, \ldots, d\}$.
La loi P est un mélange de deux lois binomiales de paramètres (d, q_0) et (d, q_1) avec $q_0 < q_1$ dans des proportions $\pi, 1-\pi$ avec $\pi \in]0, 1[$.
(a) (1 point) Dans le modèle de mélange, calculer, $\mathbb{E}X, \mathbb{E}X^2, \mathbb{E}X^3$ où $X \sim P$.

Solution: Une variable aléatoire X P-distribuée, est distribuée comme $YZ_0 + (1-Y)Z_1$ où $Y \sim \text{Ber}(\pi)$, $Z_0 \sim \text{Bin}(d, q_0)$, $Z_1 \sim \text{Bin}(d, q_1)$ avec Y, Z_0, Z_1 indépendantes. $\mathbb{E}[Z_0^2] = dq_0(1+(d-1)q_0)$ et $\mathbb{E}[Z_0^3] = dq_0\,(1+(d-1)q_0(3+(d-2)q_0))$

$$
\begin{aligned}
\mathbb{E}X &= \pi dq_0 + (1-\pi)dq_1 \\
\mathbb{E}X^2 &= \text{var}(X) + (\mathbb{E}X)^2 \\
&= \mathbb{E}\left[\text{var}(X \mid Y)\right] + \text{var}\left(\mathbb{E}[X \mid Y]\right) + (\mathbb{E}X)^2 \\
&= \pi dq_0(1-q_0) + (1-\pi)dq_1(1-q_1) + \pi(1-\pi)d^2(q_1-q_0)^2 + (\pi dq_0 + (1-\pi)dq_1)^2 \\
&= \pi dq_0(1-q_0) + (1-\pi)dq_1(1-q_1) + \pi(1-\pi)d^2(q_1^2+q_0^2) + \pi^2d^2q_0^2 + (1-\pi)^2d^2q_1^2 \\
&= \pi dq_0(1-q_0) + (1-\pi)dq_1(1-q_1) + \pi d^2q_0^2 + (1-\pi)d^2q_1^2 \\
&= \pi dq_0\left((d-1)q_0+1\right) + (1-\pi)dq_1\left((d-1)q_1+1\right) \\
\mathbb{E}X^3 &= \mathbb{E}[\mathbb{E}[X^3 \mid Y]] \\
&= \pi\left(dq_0\left(1+(d-1)q_0(3+(d-2)q_0)\right)\right) + (1-\pi)\left(dq_1\left(1+(d-1)q_1(3+(d-2)q_1)\right)\right)
\end{aligned}
$$

(b) (½ point) Le modèle de mélange est-il identifiable ?

(c) (1 point) Dans le modèle de mélange proposer une suite d'estimateurs consistants de q_0, q_1, π.

(d) (1 point) Etudier l'éventuelle normalité asymptotique de l'estimateur.

(e) (1 point) Calculer l'information de Fisher dans le modèle de mélange.

(f) (1 point) Comparaison de la variance de l'estimateur et de l'inverse de l'information de Fisher (vous pouvez comparer les matrices de covariance au sens de l'ordre semi-défini positif ou vous intéresser aux variances des estimateurs de π, q_0, q_1).

(g) (1 point) Proposer un estimateur facile à calculer et dont la variance asymptotique coïncide avec l'inverse de l'information de Fisher.

4. On dispose de deux échantillons i.i.d. collectés indépendamment l'un de l'autre $X_1, \ldots, X_n$ et $Y_1, ..., Y_m$. Les X_i sont distribués selon $\mathcal{N}(\mu_1, \sigma_1^2)$, les Y_i sont distribués selon $\mathcal{N}(\mu_2, \sigma_2^2)$.

(a) (1 point) Proposer un intervalle de confiance de niveau $\alpha \in (0,1)$ pour le rapport σ_2^2/σ_1^2.

Solution: D'après le théorème de Student,

$$
(n-1)S_{n-1}^2 := \sum_{i=1}^{n}(X_i - \overline{X}_n)^2 \sim \sigma_1^2\chi_{n-1}^2 \qquad \text{et} \qquad (m-1)S_{m-1}^2 := \sum_{i=1}^{m}(Y_i - \overline{Y}_m)^2 \sim \sigma_2^2\chi_{m-1}^2 .
$$

On choisit I_α comme un intervalle de probabilité $1-\alpha$ sous la loi de Fisher à $(m-1, n-1)$ degrés de liberté. Avec probabilité $1-\alpha$

$$
\frac{\sigma_1^2}{\sigma_2^2} \in \frac{S_{n-1}^2}{S_{m-1}^2} I_\alpha
$$

(b) (1 point) Proposer un test de niveau α pour séparer $H_O : \sigma_1^2 = \sigma_2^2$ contre $H_1 : \sigma_2^2/\sigma_1^2 > 1$.

Solution: Le test peut consister à rejeter l'hypothèse nulle lorsque 1 n'appartient pas à la région de confiance de la question précédente.

Comme il s'agit d'un test dit unilatère, on peut procéder comme suit. On compare le quantile d'ordre $1-\alpha$ de la loi de Fisher à $m-1, n-1$ degrés de liberté à S_{m-1}^2/S_{n-1}^2. Si le quantile est plus grand que S_{m-1}^2/S_{n-1}^2, on ne rejette pas l'hypothèse nulle, sinon on rejette. De cette façon le test est non seulement de niveau demandé mais en plus sans biais.

(c) (1 point) Si on suppose $\sigma_1 = \sigma_2$, proposer un intervalle de confiance de niveau $1-\alpha$, pour $\mu_1 - \mu_2$.

Solution: Dans ces conditions

$$
(\overline{X}_n - \mu_1) - (\overline{Y}_m - \mu_2) \sim \mathcal{N}\left(0, \sigma_1^2\frac{n+m}{nm}\right)
$$

Cette quantité aléatoire est indépendante de

$$\sum_{i=1}^{n}\left(X_i-\overline{X}_n\right)^2+\sum_{i=1}^{m}\left(Y_i-\overline{Y}_m\right)^2\sim\sigma_1^2\chi^2_{n+m-2}$$

On note $\widehat{\sigma}_1=\frac{1}{n+m-2}\left(\sum_{i=1}^n\left(X_i-\overline{X}_n\right)^2+\sum_{i=1}^m\left(Y_i-\overline{Y}_m\right)^2\right)$. En notant $t_{\alpha/2}$ comme le quantile d'ordre $1-\alpha/2$ de la loi de Student à $n+m-2$ degrés de liberté, on peut proposer l'intervalle de confiance

$$\overline{X}_n-\overline{Y}_m\pm t_{\alpha/2}\widehat{\sigma}_1\sqrt{\frac{n+m}{nm}}$$

(d) (1 point) Si on suppose $\sigma_1=\rho\sigma_2$ (ρ connu), proposer un intervalle de confiance de niveau $1-\alpha$, pour $\mu_1-\mu_2$.

5. La loi $P_{\gamma,\mu,\sigma}$ (avec $\gamma,\sigma>0$) est la loi de $\gamma+\exp(\sigma X+\mu)$ où $X\sim\mathcal{N}(0,1)$. On dispose d'un échantillon i.i.d. $X_1,\ldots,X_n$ de $P_{\gamma,\mu,\sigma}$.

(a) (1 point) Calculer, espérance, variance, médiane de $P_{\gamma,\mu,\sigma}$.

Solution: $P_{\gamma,\mu,\sigma}$ est une translation par γ d'une loi log-normale de paramètres μ,σ. Espérance $\gamma+\exp(\mu+\sigma^2/2)$, Variance $\exp(2\mu+\sigma^2)\left(\exp(\sigma^2)-1\right)$, Médiane $\gamma+\exp(\mu)$. L'écart interquartile est

$$2\exp(\mu)\sinh(\sigma z_{3/4})\,.$$

Le paramètre γ est un paramètre de localisation particulier puisqu'il délimite le support de la loi. Les paramètres de μ et σ influent à la fois la localisation et la dispersion de la loi.

(b) (1 point) Proposer un estimateur des moments pour (γ,μ,σ), étudier sa consistance, et éventuellement sa normalité asymptotique.

Solution: On note $\overline{X}_n$ la moyenne empirique, $\widehat{s}$ l'écart type empirique et $\widehat{m}$ la médiane empirique. On estime $\exp(\sigma^2/2)$ comme la solution positive de $\frac{\overline{X}_n-\widehat{m}}{\widehat{s}}=\sqrt{x-1}/(x\sqrt{x+1})$. A partir de l'écart type, on obtient un estimateur de μ, puis de γ.
Comme la moyenne empirique, la variance empirique et la médiane empirique sont des estimateurs consistants de l'espérance, de la variance et de la médiane, en invoquant la continuité des estimateurs par rapport à ces quantités et le principe de l'image continue, on obtient la consistance. La normalité asymptotique s'obtient par la méthode δ.

(c) (2 points) La maximisation de la vraisemblance fournit elle une méthode d'estimation consistante dans ce modèle ?

Question:	1	2	3	4	5	Total
Points:	6	6	6½	4	4	26½
Score:						