LSTAT2040: Examen

Durée 3 heures

Les feuilles blanches sont vos feuilles de réponses. Inscrivez vos nom et prénom sur chacune d'entre elles et numérotez-les dans l'ordre de lecture. Seules ces feuilles seront corrigées. Les feuilles de couleur vous serviront de brouillon.
À la fin de l'examen, vous devez rendre toutes les feuilles que vous avez reçues, y compris les feuilles de brouillon et les feuilles vierges que vous n'avez pas utilisées.
Tout échange d'informations, sous quelque forme que ce soit, est interdit et sera considéré comme une tricherie.
Toute réponse non justifiée sera considérée comme étant erronée.
Vous pouvez consulter le syllabus du cours.
L'examen est sur 22 et compte 2 points de bonus.

Excercice 1 (15 pts)

X est une variable aléatoire dont la fonction de densité est donnée par

$$f(x) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1} I(0 < x < 1),\ \ \alpha, \beta > 0,$$

où $B(\alpha,\beta) := \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$, $\forall \alpha, \beta$.
Γ est la fonction gamma $\Gamma(x) = \int_0^\infty t^{x-1}\exp(-t)\,dt$. Cette fonction possède un certain nombre de propriétés, notamment le fait que $\Gamma(x+1) = x\Gamma(x)$, $\forall x > 0$ et $\Gamma(x+1) = x!$, $\forall x \in \mathbb{N}$.
Dans ce qui suit, nous aurons également besoin de la fonction ψ (connue sous le nom `digamma`), définie comme suit

$$\psi(x) := \frac{d\log(\Gamma(x))}{dx} = \frac{\Gamma'(x)}{\Gamma(x)}.$$

Dans cet exercice, `digamma` est à considérer comme étant une fonction connue (au même titre que $\log(\cdot)$, ou $\exp(\cdot)$, par exemple), même si il n'est pas possible de la définir explicitement. L'une des caractéristiques intéressantes de cette fonction est le fait que $\psi(x+1) = \frac{1}{x} + \psi(x)$.
Sauf indication contraire, nous supposons que α et β sont inconnus et que nous observons X_i, $i = 1, \ldots, n$, un échantillon *iid* de X.

1 (2 pts)

Montrez que

$$E(X) = \frac{\alpha}{\alpha+\beta}, \text{ et } E(X^2) = \frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)}.$$

Solution

Par définition,

$$E(X^k) = \frac{1}{B(\alpha,\beta)} \int_0^1 x^{\alpha+k-1}(1-x)^{\beta-1}dx = \frac{B(\alpha+k,\beta)}{B(\alpha,\beta)}.$$

Il en résulte que,

$$\begin{aligned} E(X) &= \frac{B(\alpha+1,\beta)}{B(\alpha,\beta)} \\ &= \frac{\Gamma(\alpha+1)\Gamma(\beta)/\Gamma(\alpha+\beta+1)}{\Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha+\beta)} \\ &= \frac{\alpha\Gamma(\alpha)\Gamma(\beta)/(\alpha+\beta)\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha+\beta)} \\ &= \frac{\alpha}{\alpha+\beta}. \end{aligned}$$

En procédant de la même manière, on peut établir l'expression de $E(X^2)$ comme requis.

□

2 (2 pts)

Montrez que f fait partie des familles exponentielles. En suivant les notations du cours pour les familles exponentielles, précisez ses différentes composantes (h, B, ...) et donnez sa forme canonique. Servez-vous en pour prouver que

$$E(\log(X)) = \psi(\alpha) - \psi(\alpha+\beta), \text{ et que } Var(\log(X)) = \psi'(\alpha) - \psi'(\alpha+\beta).$$

Solution

Il est clair que nous pouvons exprimer f comme suit

$$\begin{aligned} f(x) &= \frac{1}{x(x-1)} \exp\left(\alpha\log(x) + \beta\log(1-x) - \log(B(\alpha,\beta))\right) \\ &\equiv h(x)\exp(\alpha T_1 + \beta T_2 - A(\alpha,\beta)). \end{aligned}$$

Ce qui prouve bien que f fait partie des familles exponentielles à deux paramètres. La forme "originale" et la forme canonique coïncident.

Selon les propriétés des familles exponentielles (voir cours),

$$
\begin{aligned}
&E(\log(X)) = \partial_\alpha \log B(\alpha,\beta) = \partial_\alpha \Gamma(\alpha) - \partial_\alpha \Gamma(\alpha+\beta) = \psi(\alpha) - \psi(\alpha+\beta) \\
&Var(\log(X)) = \partial_\alpha E(\log(X)) = \psi'(\alpha) - \psi'(\alpha+\beta).
\end{aligned}
$$

□

3 (2 pts)

Supposons que β soit connu. Trouver un estimateur de α par la méthode de moments (MM). Quelle est sa distribution asymptotique ? Donnez sa variance asymptotique sous forme d'une expression explicite en fonction de α et β uniquement.

Solution

Soit $\mu = E(X)$. Nous savons que $\mu = \alpha/(\alpha+\beta)$, ou, de manière équivalente, $\alpha = \beta(1/\mu - 1)^{-1}$. Cela suggère comme estimateur MM

$$
\hat{\mu}_M = \beta \frac{1}{\frac{1}{\overline{X}_n} - 1}.
$$

Soit $\sigma^2 = Var(X)$. En appliquant CLT + Delta-method, nous avons que

$$
\begin{aligned}
&\sqrt{n}(g(\overline{X}_n) - g(\mu)) \xrightarrow{d} g^{'}(\mu) N(0,\sigma^2) \\
\Longleftrightarrow\ &\sqrt{n}(\hat{\mu}_M - \alpha) \xrightarrow{d} N\left(0, \sigma^2 (g^{'}(\mu))^2\right) \\
\Longleftrightarrow\ &\hat{\mu}_M \sim_a N(\alpha, \sigma_M^2/n),
\end{aligned}
$$

avec

$$
\sigma_M^2 = \sigma^2 (g^{'}(\mu))^2 = \sigma^2 \left(\frac{\beta}{(1-\mu)^2}\right)^2 = \sigma^2 \beta^2 \left(\frac{1}{1-\mu}\right)^4,
$$

où

$$
\sigma^2 = E(X^2) - (E(X))^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)},
$$

et donc,

$$
\sigma_M^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \beta^2 \left(\frac{\alpha+\beta}{\beta}\right)^4 = \frac{\alpha(\alpha+\beta)^2}{\beta(\alpha+\beta+1)}.
$$

□

4 (3 pts)

Montrez que l'estimateur MM de $\boldsymbol{\theta} := (\alpha,\beta)$ est donné par $\hat{\boldsymbol{\theta}}_{MM} = (\hat{\alpha}_{MM}, \hat{\beta}_{MM})$, où

$$
\hat{\alpha}_{MM} = \overline{X} \frac{\overline{X} - \overline{X^2}}{\overline{X^2} - \overline{X}^2}, \text{ et } \hat{\beta}_{MM} = (1-\overline{X}) \frac{\overline{X} - \overline{X^2}}{\overline{X^2} - \overline{X}^2}.
$$

$\hat{\boldsymbol{\theta}}_{MM}$ est-il consistant ? Converge-t-il en distribution vers $\boldsymbol{\theta}$?

Solution

Soit $\mu_k = E(X^K)$. Nous avons que,

$$
\begin{aligned}
\mu_1 &= \frac{\alpha}{\alpha+\beta}, \text{ et} \\
\mu_2 &= \frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)} \\
&= \mu_1 \frac{\alpha+1}{\alpha+\beta+1} \\
&= \mu_1 \frac{1+1/\alpha}{(\alpha+\beta)/\alpha+1/\alpha} \\
&= \mu_1 \frac{1+1/\alpha}{1/\mu_1+1/\alpha} \\
\iff \frac{\mu_2}{\mu_1} &= \frac{1+1/\alpha}{1/\mu_1+1/\alpha} \\
\iff \alpha &= \frac{1-\mu_2/\mu_1}{\mu_2/\mu_1^2-1} = \mu_1 \frac{\mu_1-\mu_2}{\mu_2-\mu_1^2}.
\end{aligned}
$$

Ainsi, l'estimateur MM de α est

$$
\hat{\alpha}_{MM} = \frac{\overline{X}(\overline{X}-\overline{X^2})}{\overline{X^2}-\overline{X}^2}
$$

Par WLLN, nous savons que $(\overline{X}, \overline{X^2}) \xrightarrow{p} (\mu_1, \mu_2)$. En appliquant CMT, il est facile d'en déduire que $\hat{\mu}_{MM} \xrightarrow{p} \alpha$, càd que $\hat{\mu}_{MM}$ est consistant.

Pour β, nous avons que

$$
\mu_1 = \frac{\alpha}{\alpha+\beta} \iff \mu_1 = \frac{1}{1+\beta/\alpha} \iff \beta = \alpha(\frac{1}{\mu_1}-1) = (1-\mu_1)\frac{\mu_1-\mu_2}{\mu_2-\mu_1^2}
$$

Pour le reste, on applique le même raisonnement que pour α.

Comme conséquence, nous pouvons conclure que $\boldsymbol{\theta}$ est consistant, càd $\hat{\boldsymbol{\theta}}_{MM} \xrightarrow{p} \boldsymbol{\theta}$. Et comme la convergence en prob. implique celle en dist., $\hat{\boldsymbol{\theta}}_{MM} \xrightarrow{d} \boldsymbol{\theta}$.

□

5 (2 pts)

Donnez, en fonction de $\psi(\cdot)$, les équations de vraisemblances qui permettent de calculer l'estimateur du maximum de vraisemblance (MV) de $\boldsymbol{\theta}$. Expliquez, le plus précisément possible, la procédure à suivre pour résoudre ces équations numériquement.

Solution

On que

$$
\begin{aligned}
\ell_n &= -n\log\Gamma(\alpha) - n\log\Gamma(\beta) + n\log\Gamma(\alpha+\beta) + (\alpha-1)\sum_i \log(x_i) + (\beta-1)\sum_i \log(1-x_i)\\
S_1 &:= \partial_\alpha \ell_n = n(\psi(\alpha+\beta) - \psi(\alpha) + n^{-1}\sum_i \log(x_i))\\
S_2 &:= \partial_\beta \ell_n = n(\psi(\alpha+\beta) - \psi(\beta) + n^{-1}\sum_i \log(1-x_i))\\
H_{1,1} &:= \partial_\alpha S_1 = n(\psi'(\alpha+\beta) - \psi'(\alpha))\\
H_{1,2} &:= \partial_\beta S_1 = n\psi'(\alpha+\beta) = H_{2,1} := \partial_\alpha S_2\\
H_{2,2} &:= \partial_\beta S_2 = n(\psi'(\alpha+\beta) - \psi'(\beta)).
\end{aligned}
$$

Les équations de vraisemblance sont $S_1 = 0$ et $S_2 = 0$, càd

$$
\begin{aligned}
\psi(\alpha) - \psi(\alpha+\beta) &= n^{-1}\sum_i \log(X_i), \text{ et}\\
\psi(\beta) - \psi(\alpha+\beta) &= n^{-1}\sum_i \log(1-X_i).
\end{aligned}
$$

Pour résoudre ces équations, nous pouvons utiliser la méthode de Newton-Raphson (NR), en itérant selon la formule suivante

$$\boldsymbol{\theta}_{k+1}^t = \boldsymbol{\theta}_k^t - \boldsymbol{H}^{-1}(\boldsymbol{\theta}_k)\boldsymbol{S}(\boldsymbol{\theta}_k),\ k = 0, \ldots$$

où $\boldsymbol{S} = (S_1, S_2)^t$ et $\boldsymbol{H}$ est la Hessien dont les éléments sont données ci-dessus. Pour démarrer l'algorithme ($k = 0$), nous pouvons utiliser l'estimateur MM définit ci-dessus..

□

6 (3 pts)

Supposons que $\beta = 1$. Trouver l'expression explicite de l'estimateur MV de α. Cet estimateur est-il efficace ? Prouvez sa consistance et trouvez sa variance asymptotique.

Solution

Dans le cas où $\beta = 1$, $B(\alpha, 1) = 1/\alpha$ et on que

$$
\begin{aligned}
\ell_n &= \sum_i \log(\alpha) + (\alpha-1)\log(x_i) = n\log(\alpha) + (\alpha-1)\sum_i \log(x_i)\\
S_1 &:= \partial_\alpha \ell_n = \frac{n}{\alpha} + \sum_i \log(x_i)\\
H_{1,1} &:= \partial_\alpha S_1 = -\frac{n}{\alpha^2} < 0.
\end{aligned}
$$

Il s'en suit que l'estimateur MV de α est la racine de l'équation $S_1 = 0$, càd

$$\hat{\alpha}_V = -\frac{n}{\sum_i \log(X_i)}.$$

Cet estimateur est consistant puisque, par WLLN $n^{-1}\sum_i \log(X_i) \xrightarrow{p} E(\log(X)) = \psi(\alpha) - \psi(\alpha+1) = -\frac{1}{\alpha}$. Il suffit d'appliquer la CMT pour conclure.

Selon la théorie asymptotique des estimateurs MV,

$$\hat{\alpha}_V \sim_a N(\alpha, I_n^{-1}(\alpha)),$$

avec $I_n(\alpha) = -EH_{1,1} = \frac{n}{\alpha^2}$. Donc la variance asymptotique de $\hat{\alpha}_V$ est $\frac{\alpha^2}{n}$.

Selon le théorème de CRLB-Attainment, pour qu'un estimateur efficace existe, la quantité

$$\alpha + I_n^{-1}(\alpha)S_1 = 2\alpha + \frac{\alpha^2}{n}\sum_i \log(X_i)$$

ne doit pas dépendre de α, mais ce n'est manifestement pas le cas. Donc $\hat{\alpha}_V$ ne peut pas être efficace.

□

7 (1 pt)

Lequel des estimateurs MM ou MV choisirez-vous si vous savez que $\beta = 1$ et que votre taille d'échantillon est "grande" ?

Solution

À taille taille d'échantillon finie, il est difficile de dire lequel est le meilleur. Asymptotiquement, les deux sont non biaisés, mais l'estimateur MV est plus efficace. En effet,

$$AVar(\hat{\alpha}_M) = n^{-1}\frac{\alpha(\alpha+1)^2}{\alpha+2} > n^{-1}\alpha^2 = AVar(\hat{\alpha}_{MV}).$$

□

Excercice 2 (7 pts)

Soit X_i, $i = 1, \ldots, n$, un échantillon *iid* de X dont la densité est donnée par

$$f(x) = \frac{1}{2\theta} I(0 \leq x \leq 2\theta),$$

où $\theta > 0$ est inconnu.

1 (2 pts)

Calculez la moyenne, la variance et la fonction de répartition (cdf) de X.

Solution

En effectuant de simples calculs d'intégrales, il est facile de vérifier que

$$E(X) = \theta,\ Var(X) = \theta^2/3,\ \text{ et } F_X(x) = \frac{1}{2\theta} x I(0 \leq x \leq 2\theta) + I(x > 2\theta).$$

□

2 (2 pts)

Soit T l'estimateur du maximum de vraisemblance de θ. Trouvez T et montrez que sa densité est donnée par

$$t \mapsto \frac{n}{\theta^n} t^{n-1} I(0 \leq t \leq \theta).$$

On en déduit que T est biaisé. Effectuez une correction pour le transformer en un estimateur non-biaisé.

Solution

Il'est facile de vérifier que $T = 2^{-1} \max_i X_i$. Pour trouvez sa densité, il suffit de dériver sa cdf qui est

$$P(T \leq t) = P(max_i X_i \leq 2t) = \prod_i P(X_i \leq 2t) = (F_X(2t))^n = \left(\frac{t}{\theta}\right)^n,\ \text{pour } 0 \leq t \leq \theta$$

Il s'en suit que

$$E(T) = \frac{n}{n+1}\theta$$

et donc, T est effectivement biaisé. Et pour corriger cela, il suffit de prendre

$$\hat{\theta}_1 = \frac{n+1}{n} T.$$

□

3 (2 pts)

Soit $\hat{\theta}_1 = \frac{n+1}{2n} \max_{1 \leq i \leq n} X_i$ un estimateur de θ. Montrez que la variance (exact) de $\hat{\theta}_1$ est donnée par

$$\frac{1}{n(n+2)} \theta^2.$$

Montez que $\hat{\theta}_1$ est consistant.

Solution

$$
\begin{aligned}
Var(\hat{\theta}) &= \frac{(n+1)^2}{n^2} Var(T) \\
&= \frac{(n+1)^2}{n^2} [E(T^2) - E(T)^2] \\
&= \cdots \\
&= \frac{\theta^2}{n(n+2)}
\end{aligned}
$$

On a donc $MSE(\hat{\theta}) = \frac{1}{n(n+2)}\theta^2 \to 0$, ce qui implique la consistance de $\hat{\theta}$.

□

4 (1 pt)

Un autre estimateur possible de θ est $\hat{\theta}_2 = \overline{\mathrm{X}}$. De ces deux estimateurs ($\hat{\theta}_1$ and $\hat{\theta}_2$), lequel est le meilleur ?

Solution

$MSE(\hat{\theta}_2) = Var(\overline{\mathrm{X}}) = \frac{var(X)}{n} = \frac{\theta^2}{3n}$. Et il est facile de voir que

$$MSE(\hat{\theta}_2) > MSE(\hat{\theta}_1),\ \forall n > 1.$$

Donc $\hat{\theta}_1$ est plus efficace.

□