Modélisation statistique (STA221)
Examen (3h)

Olivier Rioul

11 octobre 2024

*Contrôle avec **notes de cours manuscrites, UNIQUEMENT.**. Calculettes autorisées.*
On détaillera avec précision tous les raisonnements.
*Rédiger lisiblement, **en faisant des phrases** en langue humaine SVP.*
Le soin et la présentation comptent pour la note finale.

I Non existence d'un estimateur non biaisé

ON CONSIDÈRE le modèle statistique i.i.d. $\mathscr{B}(\frac{1}{\theta})$ Bernoulli de paramètre $1/\theta$, où $\theta > 1$, avec un N-échantillon $X = (X_1, X_2, \dots, X_N)$.

1. Justifier que c'est un modèle régulier et trouver la borne de Cramér-Rao pour un estimateur $\hat{\theta}$ non biaisé.
2. Démontrer que $S = X_1 + X_2 + \cdots + X_N$ est une statistique suffisante. Quelle est sa loi ?

Dorénavant on s'intéresse uniquement aux estimateurs fonctions de S, de la forme $\hat{\theta} = \hat{\theta}(S)$.

3. En utilisant la loi de S, trouver une expression du *biais* d'un estimateur quelconque $\hat{\theta} = \hat{\theta}(S)$ en fonction notamment de $\theta > 1$.
4. La borne de Cramér-Rao établie ci-dessus s'applique-t-elle à $\hat{\theta}$? Pourquoi ?

II Non existence d'un estimateur MVU

ON SE DONNE un échantillon normal de deux mesures indépendantes seulement : $X = (X_1, X_2)$ avec $X_1 \sim \mathcal{N}(\theta, 1)$ pour tout $\theta \in \mathbb{R}$, et $X_2 \sim \mathcal{N}(\theta, 1)$ si $\theta > 0$, mais $X_2 \sim \mathcal{N}(\theta, 2)$ si $\theta \leqslant 0$.

1. Calculer l'information de Fisher en fonction de θ, en distinguant les cas $\theta > 0$ et $\theta \leqslant 0$.
2. Existe-t-il un estimateur efficace ?
3. Calculer le biais et la variance des deux estimateurs $\hat{\theta}_1 = \frac{X_1+X_2}{2}$ et $\hat{\theta}_2 = \frac{2X_1+X_2}{3}$ et comparer à la borne de Cramér-Rao.
4. Montrer qu'un estimateur MVU serait nécessairement efficace. Conclure.

III Estimateurs inadmissibles

ON CONSIDÈRE l'estimateur de la moyenne $\bar{X}_N$ qui estime la moyenne $\theta = \mathbf{E}(X_i)$ pour un N échantillon i.i.d. $X = (X_1, X_2, \ldots, X_N)$ avec $N > 1$. On dit qu'un estimateur est uniformément meilleur qu'un autre si son risque quadratique est toujours plus faible (strictement pour au moins une valeur du paramètre). On dit qu'un estimateur est *inadmissible* s'il existe un estimateur uniformément meilleur, *admissible* dans le cas contraire.

1. Donner le biais et la variance de $\bar{X}_N$. En déduire que l'estimateur X_1 est *inadmissible*.

On suppose dorénavant un modèle gaussien i.i.d. $\mathcal{N}(\theta, \sigma^2)$ avec $\theta \in [0,1]$.

2. Construire l'estimateur MAP noté $\hat{\theta}_N$, pour un a priori uniforme (de Laplace) sur $\theta \in [0,1]$.
3. En déduire que $\bar{X}_N$ est *inadmissible* pour ce modèle.
4. Montrer que l'estimateur constant $\hat{\theta} = 1/2$ est *admissible*. Commenter.

IV Règle de succession de Laplace

ON CONSIDÈRE le modèle bayesien i.i.d. $\mathscr{B}(\theta)$ Bernoulli de paramètre θ, dont on rappelle qu'une statistique suffisante est $S = X_1 + X_2 + \cdots + X_N$, avec un a priori uniforme (de Laplace) pour $\theta \in [0,1]$.

1. Trouver la loi de l'a posteriori, en fonction de S.
2. En déduire l'estimateur MMSE $\hat{\theta}(X)$ et le comparer au ML.
3. Démontrer que $\hat{\theta}(X) = \mathbb{P}(X_{N+1} = 1 | X_1 + X_2 + \cdots + X_N = S)$.
4. Sachant que le soleil s'est levé tous les jours "depuis le début de la Terre" (soit 5000 ans selon Laplace, 1814), quelle la probabilité que le soleil se lève demain ?

Solutions

I-1) Pour un échantillon, $p_\theta(x) = (1/\theta)^x(1-1/\theta)^{1-x}$ de support $\{0,1\}$ et différentiable en θ. Le modèle est donc régulier. On a $\log p_\theta(x) = -x\log(\theta)+(1-x)(\log(\theta-1)-\log\theta)$, score $S_\theta(X) = -\frac{X}{\theta}+(1-X)(\frac{1}{\theta-1}-\frac{1}{\theta}) = -\frac{X}{\theta}+\frac{1-X}{\theta(\theta-1)}$, information de Fisher $J_{\theta,1} = \mathbf{V}(S_\theta(X)) = (\frac{1}{\theta}+\frac{1}{\theta(\theta-1)})^2\frac{1}{\theta}(1-\frac{1}{\theta}) = \frac{1}{\theta^2(\theta-1)}$, $J_\theta = \frac{N}{\theta^2(\theta-1)}$ car le modèle est i.i.d., d'où la CRB=$\frac{\theta^2(\theta-1)}{N}$.

N.B. : On peut aussi obtenir J_θ par reparamétrisation $\theta \to 1/\theta$ à partir de $J_\theta = \frac{N}{\theta(1-\theta)}$ pour le modèle i.i.d. $\mathscr{B}(\theta)$. On obtient alors $J_\theta = \frac{N}{(1/\theta)(1-1/\theta)}/g'(1/\theta)^2$ où $g(x) = 1/x$, soit $J_\theta = \frac{N}{(\theta^4/\theta)(1-1/\theta)} = \frac{N}{\theta^2(\theta-1)}$.

I-2) On vérifie la factorisation de Fisher : $p_\theta(x) = \prod_1^N(1/\theta)^{x_i}(1-1/\theta)^{1-x_i} = (\frac{1}{\theta-1})^s(1-1/\theta)^N$ où $s = \sum_1^N x_i$, d'où le résultat. Puisque $X \sim \mathscr{B}(\frac{1}{\theta})$ i.i.d., on sait que $S = X_1+X_2+\cdots+X_N \sim \mathscr{B}(N,\frac{1}{\theta})$, loi binomiale.

I-3) $B(\hat\theta) = \mathbf{E}(\hat\theta(S))-\theta = \sum_{s=0}^N \hat\theta(s)\binom{N}{s}(\frac{1}{\theta})^s(1-\frac{1}{\theta})^{N-s}-\theta$.

I-4) Si l'estimateur $\hat\theta = \hat\theta(S)$ est non biaisé (pour tout $\theta > 1$), on obtient l'identité suivante : $\sum_{s=0}^N \hat\theta(s)\binom{N}{s}(\frac{1}{\theta})^s(1-\frac{1}{\theta})^{N-s} = \theta$ de la forme $P(1/\theta) = \theta$ où P est un polynôme de degré N ; identité qui doit être satisfaite pour tout $\theta > 1$ pour que la borne de Cramér-Rao ci-dessus s'applique sur la variance de $\hat\theta$. Mais ceci est impossible, car on aurait $P(x) = 1/x$ fraction rationnelle et non polynôme. Donc la borne de Cramér-Rao établie ci-dessus ne s'applique pas, puisqu'il n'existe *pas* d'estimateur $\hat\theta = \hat\theta(S)$ non biaisé pour tout $\theta > 1$.

II-1) Pour $\theta > 0$, on trouve comme dans le cas normal classique vu en cours, $J_\theta = N/\sigma^2 = 2/1 = 2$. Pour $\theta \leqslant 0$, on a $p_\theta(x) = \frac{1}{2\sqrt{2}\pi}\exp-[\frac{(x_1-\theta)^2}{2}+\frac{(x_2-\theta)^2}{4}]$, d'où $S_\theta(X) = (X_1-\theta)+\frac{X_2-\theta}{2}$ et $J_\theta = \mathbf{V}(S_\theta(X)) = 1+\frac{2}{4} = \frac{3}{2}$.

II-2) S'il existe, l'estimateur efficace aurait donc pour expression $\theta(X) = \theta+\frac{S_\theta(X)}{J_\theta}$ qui vaut $\frac{X_1+X_2}{2}$ si $\theta > 0$, et $\frac{2X_1+X_2}{3}$ si $\theta \leqslant 0$. Mais cette expression dépend de la valeur de θ, donc l'estimateur efficace n'existe pas.

II-3) Ces deux estimateurs sont non biaisés, de variance $\mathbf{V}(\hat\theta_1) = 1/2$, $\mathbf{V}(\hat\theta_2) = 10/18 > 1/2$ pour $\theta > 0$, et $\mathbf{V}(\hat\theta_1) = 3/4 > 2/3$, $\mathbf{V}(\hat\theta_2) = 2/3$ pour $\theta \leqslant 0$. Ainsi $\hat\theta_1$ atteint la CRB pour $\theta > 0$ (mais *pas* pour $\theta \leqslant 0$), et $\hat\theta_2$ atteint la CRB pour $\theta \leqslant 0$ (mais *pas* pour $\theta \leqslant 0$).

II-4) Puisque la CRB est atteinte à la fois pour $\theta > 0$ et pour $\theta \leqslant 0$, un estimateur MVU atteindrait nécessairement la CRB pour tout θ puisqu'il serait au moins aussi bon que $\hat\theta_1$ *et* $\hat\theta_2$. Il serait donc efficace, ce qui est impossible comme on l'a vu.

III-1) $\bar{X}_N$ est non biaisé de variance $\frac{\sigma^2}{N}$ où σ^2 est la variance d'un échantillon. De même $X_2 = \bar{X}_1$ est non biaisé de variance $\frac{\sigma^2}{1} = \sigma^2 > \frac{\sigma^2}{N}$ car $N > 1$, il est donc inadmissible puisque $\bar{X}_N$ est uniformément meilleur.

III-2) $\theta = \hat{\theta}_N$ maximise $p(x|\theta)1_{[0,1]}(\theta)$, donc minimise $\sum_1^N |X_i - \theta|^2 1_{[0,1]}(\theta)$, c'est donc $\hat{\theta}_N = \bar{X}_N$ si $0 < \theta < 1$, $\hat{\theta}_N = 1$ si $\bar{X}_N > 1$ et $\hat{\theta}_N = 0$ si $\bar{X}_N \leqslant 0$. Autrement dit $\hat{\theta}_N = \bar{X}_N 1_{[0,1]}(\bar{X}_N)$.

III-3) On voit facilement que $|\theta - \hat{\theta}_N| \leqslant |\theta - \bar{X}_N|$ avec inégalité stricte si $\bar{X}_N \leqslant 0$ ou $\bar{X}_N \geqslant 1$, d'où $\hat{\theta}_N$ est uniformément meilleur que $\bar{X}_N$ (pour le MSE).

III-4) Si un certain estimateur $\hat{\theta}'$ était meilleur, on aurait $\mathbf{E}((1/2-\theta)^2) \geqslant \mathbf{E}((\hat{\theta}' - \theta)^2)$ pour tout $\theta \in [0,1]$, d'où en faisant $\theta = 1/2$ il vient $\hat{\theta}' = \hat{\theta} = 1/2$. Il n'existe donc aucun estimateur uniformément meilleur (strictement meilleur pour au moins une valeur du paramètre). L'estimateur constant $\hat{\theta} = 1/2$ est donc admissible.

Commentaire : $\hat{\theta} = 1/2$, bien que très mauvais, est admissible, alors que l'estimateur efficace (MVU) $\bar{X}_N$ est inadmissible. Cela relativise donc l'utilité de la notion d'"estimateur admissible".

IV-1) On a $p(X|\theta) = \prod_1^N (\theta)^{X_i}(1-\theta)^{1-X_i} = \theta^S(1-\theta)^{N-S}$, puis $p(\theta|X) \propto (X|\theta)p(\theta) = \theta^S(1-\theta)^{N-S}$, c'est une loi Bêta $B(S+1, N-S+1)$.

IV-2) La moyenne d'une loi $B(\alpha,\beta)$ est $\frac{\alpha}{\alpha+\beta}$, d'où $\hat{\theta}(X) = \frac{S+1}{N+2}$. L'estimateur ML est $\bar{X}_N = \frac{S}{N}$.

IV-3) $\mathbb{P}(X_{N+1} = 1|X_1 + X_2 + \cdots + X_N = S) = \mathbf{E}(X_{N+1}|X_1 + X_2 + \cdots + X_N = S) = \mathbf{E}\mathbf{E}(X_{N+1}|\theta, S)$ où sachant θ, X_{N+1} est indépendant de S et suit la loi $\mathscr{B}(\theta)$ de moyenne θ, d'où $\mathbb{P}(X_{N+1} = 1|X_1 + X_2 + \cdots + X_N = S) = \mathbf{E}(\theta|S) = \hat{\theta}(X) = \frac{S+1}{N+2}$.

IV-4) Ici $X_i = 1$ signifie que le soleil se lève le $i$ième jour. Sachant que $S = N = 5000 \times 365 \approx 1825000$, la probabilité en question est $= \frac{N+1}{N+2} = \frac{1825001}{1825002} = 0,999999452\ldots$