Statistique mathématique – partiel

M1 Mathématiques Fondamentales, Université Paris-Saclay
2024-2025

Lundi 17 février 2025
9h – 12h

Une feuille A4 recto manuscrite est autorisée en tant que support. La calculatrice n'est pas autorisée.

Avant de commencer :

- *Le sujet comporte quatre exercices indépendants. Il est demandé de les traiter dans l'ordre sur la copie.*
- *Des résultats pourront être admis d'une question sur l'autre, à la condition de l'écrire clairement.*
- *Une attention particulière sera portée aux questions d'interprétation, à la rigueur et à la précision de la rédaction.*
- *Le sujet est long et n'est pas nécessairement pensé pour être terminé dans son intégralité dans le temps imparti.*
- *Un barème indicatif est donné :*

  Exercice 1 : 4 points,

  Exercice 2 : 8 points,

  Exercice 3 : 4 points.

  Exercice 4 : 4 points.

Bon courage !

Exercice 1 – Fonction caractéristique des variables gaussiennes

Nous voulons montrer que, comme indiqué dans le cours, si $X \sim \mathcal{N}(\mu, \sigma^2)$, alors pour tout $t \in \mathbb{R}$,

$$\Phi_X(t) = \exp\left(i\mu t - \frac{\sigma^2}{2}t^2\right).$$

0. Montrer le résultat lorsque X est dégénérée ($\sigma = 0$). *Solution. Dans ce cas, $X + \mu$ presque sûrement, et $\Phi_X(t) = \mathbb{E}[e^{iXt}] = \exp(i\mu t)$.*

   On suppose dans toute la suite que $\sigma > 0$.

1. Montrer que, pour tout $t \in \mathbb{R}$, on a

   $$\Phi_X(t) = e^{it\mu} f_\sigma(t),$$

   où

   $$f_\sigma(t) := \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{i\sigma t y - y^2/2}\, dy\,.$$

   Solution. Écrire simplement la définition de la fonction caractéristique et effectuer le changement de variables $y = \frac{x-\mu}{\sigma}$.

2. Montrer que, pour tout $t \in \mathbb{R}$, $f_\sigma(t) \in \mathbb{R}$, que f_σ est différentiable, et trouver une équation différentielle qu'elle satisfait. *Solution. Prendre le conjugué complexe et effectuer le changement de variable $y \to -y$. On peut utiliser la convergence dominée car chaque intégrande est différentiable par rapport à t, et*

   $$\left| -\sigma \sin(\sigma t y)\, y\, e^{-y^2/2} \right|$$

   est facilement majorée par une fonction intégrable. En effectuant une intégration par parties (en intégrant $ye^{-y^2/2}$), on obtient :

   $$\begin{aligned} f'_\sigma(t) &= -\frac{\sigma}{\sqrt{2\pi}} \int_{\mathbb{R}} \sin(\sigma t y)\, y\, e^{-y^2/2}\, dy \\ &= 0 + \frac{\sigma}{\sqrt{2\pi}} \int_{\mathbb{R}} \sigma t \cos(\sigma t y)\, (-e^{-y^2/2})\, dy \\ &= -\sigma^2 t\, f_\sigma(t)\,. \end{aligned}$$

3. Conclure. *Solution. Nous avons la condition initiale $f_\sigma(0) = 1$. On résout l'équation différentielle précédente, ce qui donne $f_\sigma(t) = \exp(-\sigma^2 t^2/2)$. Le résultat est ainsi démontré.*

4. En déduire que si $X \sim \mathcal{N}(\mu, \sigma^2)$, et si $a, b \in \mathbb{R}$, alors

   $$aX + b \sim \mathcal{N}(a\mu + b, a^2\sigma^2).$$

   Solution. Il suffit d'écrire les fonctions caractéristiques et de reconnaître le forme souhaitée.

5. En déduire que si $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$, $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$ et que X_1 et X_2 sont indépendantes, alors

   $$X_1 + X_2 \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)\,.$$

   Solution. Il suffit d'écrire les fonctions caractéristiques et de reconnaître le forme souhaitée.

Exercice 2 – Estimation dans un modèle quasi-géométrique

On considère la variable aléatoire X prenant des valeurs dans $\mathbb{N}\cup\{-1\} = \{-1, 0, 1, 2, \dots\}$ avec la distribution suivante, paramétrée par $p \in]0, 1[$:

$$\mathbb{P}_p(X = -1) = p, \quad \text{et} \quad \forall k \geq 0,\ \mathbb{P}_p(X = k) = (1-p)^2 p^k\,.$$

Nous considérons un modèle où $(X_1, \dots, X_n)$ sont n variables aléatoires indépendantes de même loi que X.

Solution. Dans tout cet exercice, on rappelle les résultats classiques suivants. Pour $p \in]0, 1[$,

$$\sum_{k\geq 0} k p^k = p \sum_{k\geq 0} k p^{k-1} = \frac{p}{(1-p)^2}$$

et

$$\sum_{k\geq 0} k^2 p^k = p^2 \sum_{k\geq 0} k(k-1)p^{k-2} + p\sum_{k\geq 0} k p^{k-1} = \frac{2p^2}{(1-p)^3} + \frac{p}{(1-p)^2} = \frac{p(p+1)}{(1-p)^3}\,.$$

Suffisance dans le modèle

1. Calculer $\mathbb{E}_p[X_1]$ pour tout $p \in]0, 1[$. *Solution. On a*

$$\mathbb{E}_p[X_1] = -p + \sum_{k\geq 0} k(1-p)^2 p^k = -p + (1-p)^2 \frac{p}{(1-p)^2} = 0\,.$$

2. Montrer que la vraisemblance du modèle se met sous la forme

$$L(p, x_1, \dots, x_n) = (1-p)^{2n} \left(\frac{p}{(1-p)^2}\right)^{nU(x_1,\dots,x_n)} p^{nV(x_1,\dots,x_n)},$$

où

$$U(x_1, \dots, x_n) := \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{x_i=-1} \quad \text{et} \quad V(x_1, \dots, x_n) := \frac{1}{n}\sum_{i=1}^n x_i \mathbb{1}_{x_i\geq 0}\,.$$

Solution. On écrit

$$\begin{aligned} L(p, x_1, \dots, x_n) &= \prod_{i=1}^n p^{\mathbb{1}_{x_i=-1}} [(1-p)^2 p^{x_i}]^{\mathbb{1}_{x_i\geq 0}} \\ &= p^{\sum_{i=1}^n \mathbb{1}_{x_i=-1}} (1-p)^{2\sum_{i=1}^n \mathbb{1}_{x_i\geq 0}} p^{\sum_{i=1}^n x_i \mathbb{1}_{x_i\geq 0}} \\ &= p^{nU}(1-p)^{2(n-nU)} p^{nV} = (1-p)^{2n}\left(\frac{p}{(1-p)^2}\right)^{nU} p^{nV}\,. \end{aligned}$$

3. En déduire une statistique suffisante du modèle. *Solution. La densité jointe en $x = ((x_1, \dots, x_n)$ s'écrit d'après la question 2, sous la forme $h(x) g_P(U(x), V(x))$ avec $h : x \mapsto 1$ mesurable (constante) et $g_p(u, v) = (1-p)^{2n}\left(\frac{p}{(1-p)^2}\right)^{nu} p^{nv}$ mesurable. D'après le Théorème de Neyman-Fisher, (U, V) est une statistique suffisante.*

4. Cette statistique est-elle complète ? Justifier. *Solution. On a établi que les variables sont centrées dans la question 1. Mais $V - U$ n'est autre que la moyenne empirique.*

En effet,

$$V - U = \frac{1}{n}\sum_{i=1}^{n}(x_i \mathbb{1}_{x_i \geq 0} - \mathbb{1}_{x_i = -1}) = \frac{1}{n}\sum_{i=1}^{n} x_i \,,$$

elle vérifie donc $\mathbb{E}_p[U - V] = n\mathbb{E}_p[X_1] = 0$ *pour tout* $p \in]0,1[$*, alors que* $(u,v) \mapsto u - v$ *n'est jamais nulle* $\mathbb{P}_p-$*presque partout. On en déduit que* $(U(X), V(X))$ *n'est pas complète.*

Estimation par maximum de vraisemblance

5. Montrer que pour cette vraisemblance, il existe un unique estimateur du maximum de vraisemblance, noté $\hat{p}$, dont on donnera l'expression. *Solution. Comme* $p \in]0,1[$*,* L *est strictement positive, et on travaille donc avec* $\ell = \log L$*, qui s'écrit, en notant* $x = (x_1, \ldots, x_n)$*:*

$$\begin{aligned}\ell(p,x) &= 2n\log(1-p) + nU(x)\log(p) - 2nU(x)\log(1-p) + nV(x)\log(p) \\ &= n[2(1-U(x))\log(1-p) + (U(x)+V(x))\log(p)]\,.\end{aligned}$$

Cette fonction est dérivable en p *sur* $]0,1[$*, et*

$$\ell'(p,x) = n\left[-\frac{2(1-U(x))}{1-p} + \frac{U(x)+V(x)}{p}\right].$$

On résout $\ell'(p,x) > 0$ *qui est équivalent à* $(1-p)(U(x)+V(x)) > 2p(1-U(x))$ *c'est à dire* $p < \frac{U(x)+V(x)}{2+V(x)-U(x)} := p^*$ *(le dénominateur n'est jamais nul car* $2 + V(x) - U(x) \geq 2 + 0 - 1 = 1$*). La fonction* $p \mapsto \ell(p,x)$ *est donc strictement croissante sur* $]0,p^*[$ *(éventuellement vide, si* $U(x) = V(x) = 0$*) puis strictement décroissante sur* $]p^*,1[$ *(éventuellement vide, si* $U(x) = 1$ *et* $V(x) = 0$*). Elle admet un unique maximum global sur* $]0,1[$*, l'estimateur du maximum de vraisemble existe et est unique, donné par*

$$\hat{p}(X) := \frac{U(X)+V(X)}{2+V(X)-U(X)}\,.$$

On peut vérifier que $\hat{p}(X)$ *est toujours dans* $[0,1]$*: la positivité découle de celle de* U *et* V *et du fait que* $2 + V(X) - U(X) \geq 2 + 0 - 1 = 1$*, et* $\hat{p}(X) \leq 1 \iff U(X) \leq 1$ *ce qui est vrai.*

6. Montrer que $\hat{p}$ est fortement consistant. *Solution.* $\mathbb{1}_{X_1=-1}$ *est d'espérance finie* p*, et* $X_1\mathbb{1}_{X_1\geq 0}$ *est d'espérance finie*

$$\sum_{k\geq 0} k(1-p)^2 p^k = (1-p)^2 \frac{p}{(1-p)^2} = p\,.$$

Comme les variables $X_1, \ldots, X_n$ *sont i.i.d., on peut appliquer la loi forte des grands nombres à* U *et à* V *qui donne les convergences* $U \xrightarrow[n\to\infty]{p.s.} p$ *et* $V \xrightarrow[n\to\infty]{p.s.} p$*. Comme la fonction* $(u,v) \mapsto \frac{u+v}{2+v-u}$ *est continue sur* $[0,1] \times \mathbb{R}_+$*, on a* $\hat{p}(X) \xrightarrow[n\to\infty]{p.s.} \frac{p+p}{2+p-p} = p$*. L'estimateur* $\hat{p}$ *est fortement consistant.*

7. Montrer que $\hat{p}$ est asymptotiquement normal. On donnera les paramètres de la loi normale correspondante en fonction de p. *On pourra d'abord considérer la statistique*

$$S = \frac{1}{2}(U+V)$$

et montrer qu'elle est asymptotiquement normale. *Solution. Suivons l'indication et remarquons que S s'écrit $S(X) = \frac{1}{2n}\sum_{i=1}^n (\mathbb{1}_{X_i=-1} + X_i\mathbb{1}_{X_i\geq 0}) = \frac{1}{n}|X_i|/2$. Les X_i étant i.i.d. on peut appliquer le théorème central limite, sous réserve que $|X_1|$ soit bien dans L^2. Vérifions cela par le calcul. Pour $p \in]0,1[$,*

$$\mathbb{E}_p[|X_1|] = p + \sum_{k\geq 0} k(1-p)^2 p^k = 2p,$$

d'où $\mathbb{E}_p[|X_1|/2] = p$, sans surprise. De plus

$$\mathbb{E}_p[|X_1|^2] = \mathbb{E}_p[X_1^2] = p + \sum_{k\geq 0} k^2(1-p)^2 p^k = p + \frac{p(p+1)}{1-p},$$

d'où $\mathrm{Var}_p[|X_1|/2] = \frac{1}{4}\left(p + \frac{p(p+1)}{1-p} - 4p^2\right) = \frac{p(1-2p+2p^2)}{2(1-p)}$. On a donc, d'après le théorème central limite :

$$\sqrt{n}(S-p) \xrightarrow[n\to\infty]{(d)} \mathcal{N}\left(0, \frac{p(1-2p+2p^2)}{2(1-p)}\right).$$

Puis, comme $U-V$ converge p.s. donc en probabilité vers 0, on aurati envie d'appliquer le Lemme de Slutsky pour établir que

$$\sqrt{n}(\hat{p}-p) = \sqrt{n}\left(\frac{2}{2+V-U}S - p\right) \xrightarrow[n\to\infty]{(d)} \mathcal{N}\left(0, \frac{p(1-2p+2p^2)}{2(1-p)}\right).$$

Mais il faut être prudent. En effet,

$$\sqrt{n}(\hat{p}-p) = \sqrt{n}\left(\frac{2}{2+V-U} - 1\right) + \sqrt{n}\,(S-p)$$

et rien ne nous dit que $\sqrt{n}\left(\frac{2}{2+V-U} - 1\right)$ tende en loi vers 0. A vrai dire, c'est même faux, car on 'zoome' à l'échelle $1/\sqrt{n}$.
La solution demandait un peu plus de travail, est était aux frontières du programme vu ensemble. Une bonne façon de faire est d'établir la normalité asymptotique du vecteur aléatoire $(V+U, V-U)$. Cela s'établit apr le théorème central limite multidimensionnel qui s'écrit ici

$$\sqrt{n}\left((V+U, V-U) - (2p, 0)\right) \xrightarrow[n\to\infty]{(d)} \mathcal{N}\left((0,0), \Sigma = \frac{2}{1-p}\begin{pmatrix} 1-2p+2p^2 & p^2 \\ p^2 & p \end{pmatrix}\right),$$

les coefficients de Σ sont obtenus grâce aux calculs suivants :

$$\mathrm{Var}_\theta(X_1\mathbb{1}_{X_1\geq 0} + \mathbb{1}_{X_1=-1}) = \mathrm{Var}_\theta(|X_1|) = \frac{2(1-2p+2p^2)}{1-p},$$

$$\mathrm{Var}_\theta(X_1\mathbb{1}_{X_1\geq 0} - \mathbb{1}_{X_1=-1}) = \mathrm{Var}_\theta(X_1) = p + \frac{p(p+1)}{1-p} = \frac{2p}{1-p},$$

$$\mathbb{E}_\theta[(X_1\mathbb{1}_{X_1\geq 0} - \mathbb{1}_{X_1=-1})(X_1\mathbb{1}_{X_1\geq 0} + \mathbb{1}_{X_1=-1})] = \mathbb{E}_\theta[X_1^2\mathbb{1}_{X_1\geq 0}] - p = \frac{p(p+1)}{1-p} - p = \frac{2p^2}{1-p}.$$

On applique ensuite la méthode delta multidimensionnelle à $g : (x,y) \mapsto \frac{x}{2+y}$ de gradient $\nabla_g(x,y) = (\frac{1}{2+y}, -\frac{x}{(2+y)^2})$, qui, évalué en $(2p,0)$ donne $\nabla_g(2p,0) = (\frac{1}{2}, -\frac{p}{2})$. On a

$$\sqrt{n}\,(\hat{p}-p) \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0, v(p)),$$

avec $v(p) = (\frac{1}{2}, -\frac{p}{2})\Sigma(\frac{1}{2}, -\frac{p}{2})^T = \frac{1}{1-p}(\frac{1}{2}, -\frac{p}{2})(1 - 2p + 2p^2 - p^3, 0)^T = \frac{1-2p+2p^2-p^3}{2(1-p)} = \frac{1-p(1-p)}{2}$.

Cas d'une observation : estimation UVMB

Dans tout le reste de l'exercice, nous nous focalisons sur le cas d'une seule observation ($n = 1$), et l'on note $X_1 = X$ pour alléger les notations.

8. Montrer que l'ensemble des estimateurs $T : \mathbb{N} \cup \{-1\} \to \mathbb{R}$ sans biais de p est un espace affine de dimension 1, que vous décrirez. *On pourra commencer par caractériser l'ensemble*

$$\mathcal{U}_0 := \{T : \mathbb{N} \cup \{-1\} \to \mathbb{R} \,|\, \forall p \in]0,1[, \mathbb{E}_p[T(X)] = 0\}\,.$$

Solution. Suivons l'indication et prenons $U \in \mathcal{U}_0$. On a donc pour tout $p \in]0,1[$, $\mathbb{E}_p[T(X)] = 0$, c'est à dire, en notant $\alpha = U(-1)$,

$$\begin{aligned}
& pU(-1) + (1-p)^2 \sum_{k=0}^{+\infty} U(k)p^k = 0 \\
\iff & (1-p)\sum_{k=0}^{+\infty} U(k)p^k = -\frac{p}{1-p}\alpha \\
\iff & \sum_{k=0}^{+\infty} U(k)p^k - \sum_{k=0}^{+\infty} U(k)p^{k+1} + \alpha = -\frac{p}{1-p}\alpha \\
\iff & \sum_{k=0}^{+\infty} (U(k) - U(k-1))p^k = -\frac{1}{1-p}\alpha = -\sum_{k=0}^{+\infty} \alpha p^k,
\end{aligned}$$

et par unicité des coefficients des séries entières sur le disque ouvert de convergence, on en déduit $U(k) - U(k-1) = -\alpha$ pour tout $k \geq 0$ et partant $U(k) = k\alpha$ pour tout $k \geq -1$. Autrement dit tout $U \in \mathcal{U}_0$ s'écrit $U = \alpha X$ avec $\alpha \in \mathbb{R}$. On a l'égalité de l'ensemble car $\alpha X \in \mathcal{U}_0$ pour tout $\alpha \in \mathbb{R}$ d'après la question 1. Comme l'ensemble $\mathcal{U}$ des estimateurs sans biais de p est un espace affine de direction $\mathcal{U}_0$ qui est de dimension 1, et que $\mathbb{1}_{X=-1}$ est un point de cet espace, l'espace cherché s'écrit

$$\mathcal{U} = \mathbb{1}_{X=-1} + \mathrm{Vect}_{\mathbb{R}}(X)\,.$$

9. Que dire d'un estimateur UVMB de p ? *Solution. Pour chercher un estimateur UVMB de p, nous cherchons $T \in \mathcal{U}$ de variance uniformément minimale. Pour $p \in]0,1[$, cherchon $\alpha \in \mathbb{R}$ tel que $T_\alpha := \mathbb{1}_{X=-1} + \alpha X$ soit de variance minimale. On a*

$$\begin{aligned}
\mathbb{E}_p[T_\alpha^2] &= p(1-p) + 2\mathbb{E}_p[\alpha X \mathbb{1}_{X=-1}] + \alpha^2 \mathbb{E}_p[X^2] \\
&= p(1-p) - 2\alpha p + \alpha^2(p + \frac{p(p+1)}{1-p}),
\end{aligned}$$

d'où

$$\mathrm{Var}_p[T_\alpha] = p(1-p) - 2\alpha p + \alpha^2(p + \frac{p(p+1)}{1-p}) - p^2,$$

qui est minimale en $\alpha = \frac{2p}{2(p+\frac{p(p+1)}{1-p})} = \frac{1-p}{2}$. Le α optimal dépend donc de p, ainsi il n'y a pas d'estimateur de variance minimale uniformément en p. Il n'existe pas d'estimateur UVMB de p.

10. Voyez-vous un lien entre la réponse à la question 9 et celle de la question 4 ? Interpréter. *Solution. On a montré en question 4 que la statistique suffisante (U, V) n'est pas complète. Cela est compatible avec le résultat de la question 9. En fait, il n'y a pas de statistique suffisante et complète dans ce modèle, sinon, le théorème de Lehmann-Scheffé impliquerait l'exiostence (et l'unicité) d'un estimateur UVMB de p.*

Exercice 3 – Information de Fisher dans un modèle linéaire gaussien

On considère le modèle statistique où $X_1, \dots, X_n$ sont des variables réelles indépendantes, avec pour tout $1 \leq i \leq n$, $X_i \sim \mathcal{N}(\alpha + \beta t_i, 1)$, où les constantes $(t_i)_{1 \leq i \leq n}$ sont *connues* et $\alpha, \beta \in \mathbb{R}$ sont des paramètres *inconnus*.

1. Donner une condition nécessaire et suffisante sur les t_i pour que le modèle soit identifiable. Interpréter. *Solution. On observe des réalisations de Gaussiennes indépendantes de moyennes $\alpha + t_1\beta, \dots, \alpha + t_n$. Si les t_i sont constants à t, le modèle est i.i.d. de loi $\mathcal{N}(\alpha + t\beta, 1)$ et la transformation $(\alpha, \beta) \mapsto (\alpha + t\beta, 0)$ est invariante pour le modèle, il n'est donc pas identifiable. S'il existe au moins deux valeurs de t_i distinctes, mettons $t_1 \neq t_2$ alors le système formé des équations $E_1 = \alpha + t_1\beta$ et $E_2 = \alpha + t_2\beta$ est inversible, et le modèle est identifiable.*

   Dans toute la suite, on se placera sous les conditions d'identifiabilité de la question 1.

2. Déterminer la matrice d'information de Fisher $I(\alpha, \beta)$. Comment se traduit sur $I(\alpha, \beta)$ la condition de la question 1 ? Interpréter. *Solution. On écrit la log-vraisemblance qui est bien définie et vaut $\ell(\alpha, \beta, x) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^n (x_i - \alpha + t_i\beta)^2$, différentiable sur $\mathbb{R}^2$ et de gradient donné par $\nabla_{\alpha,\beta}\ell(\alpha, \beta, x) = \begin{pmatrix} \sum_{i=1}^n (x_i - \alpha + t_i\beta) \\ \sum_{i=1}^n t_i(x_i - \alpha + t_i\beta) \end{pmatrix}$. La matrice de covariance de $\nabla_{\alpha,\beta}\ell(\alpha, \beta, X)$ existe et vaut*

$$I(\alpha, \beta) = \begin{pmatrix} n & \sum_{i=1}^n t_i \\ \sum_{i=1}^n t_i & \sum_{i=1}^n t_i^2 \end{pmatrix}.$$

   A noter que l'information de Fischer est ici constante, et la matrice $I(\alpha, \beta)$ est (symétrique) et inversible si et seulement si $\det I(\alpha, \beta) \neq 0$ i.e.

$$n\sum_{i=1}^n t_i^2 \neq (\sum_{i=1}^n t_i)^2.$$

   Mais par Cauchy-Schwarz, $n\sum_{i=1}^n t_i^2 \geq (\sum_{i=1}^n t_i)^2$ avec égalité si et seulement si $(t_1, \dots, t_n)$ et $(1, \dots, 1)$ sont colinéaires c'est-à-dire si les t_i sont constants. La condition d'identifiabilité est donc la même que celle de l'inversibilité de $I(\alpha, \beta)$. C'est logique : si $I(\alpha, \beta)$ n'est pas inversible, elle l'est en aucun (α, β) (car elle est constante), et donc la log-vraisemblance parcourt localement dans un sous-espace de dimension 1. Il existe un sous-espace non trivial sur lequel la log-vraisemblance ne bouge pas quand bien même α et β bougent. Cela entraîne bien un problème d'identifiabilité.

3. Donner une borne inférieure sur la variance de tout estimateur non biaisé de α, suffisamment lisse. Vous définirez ce que 'suffisamment lisse' veut dire. *Solution. On vérifie aisément que $\mathbb{E}[\nabla_{\alpha,\beta}\ell(\alpha, \beta, X)] = 0$, la première condition du Théorème de Cramer-Rao est satisfaite. Pour l'autre condition, on doit imposer à l'estimateur T non biaisé de vérifier $\mathbb{E}[T(X)\nabla_{\alpha,\beta}\ell(\alpha, \beta, X)^T] = J_\phi(\alpha, \beta) = \begin{pmatrix} 1 & 0 \end{pmatrix}$, ici $\phi(\alpha, \beta) = \alpha$. La borne de Cramer-Rao donne qu'alors, pour de tels T et pour tout α, β,*

$$\mathrm{Var}_{\alpha,\beta}(T) \geq J_\phi(\alpha, \beta) I(\alpha, \beta)^{-1} J_\phi(\alpha, \beta)^T$$

. On a, sous les conditions d'identifiabilité,

$$I(\alpha,\beta)^{-1} = \frac{1}{n\sum_{i=1}^n t_i^2 - (\sum_{i=1}^n t_i)^2}\begin{pmatrix} \sum_{i=1}^n t_i^2 & \sum_{i=1}^n t_i \\ \sum_{i=1}^n t_i & n \end{pmatrix}.$$

ce qui donne

$$\mathrm{Var}_{\alpha,\beta}(T) \geq \frac{\sum_{i=1}^n t_i^2}{n\sum_{i=1}^n t_i^2 - (\sum_{i=1}^n t_i)^2}.$$

Cette borne est d'autant plus grande que les t_i sont grands et proches les uns des autres: c'est logique, l'estimation de α est plus difficile car la partie en $t_i\beta$ prend plus de signal.

4. Supposons que l'on connaît β. Donner dans ce cas une borne inférieure sur la variance de tout estimateur non biaisé de α suffisamment lisse. *Solution. Dans ce cas $I(\alpha) = n$ et la borne est unidimensionnelle, elle s'écrit*

$$\mathrm{Var}_{\alpha,\beta}(T) \geq \frac{1}{n}.$$

On peut remarquer que l'estimateur de la moyenne emprique, recentré par les $t_i\beta$, est alors efficace.

5. Comment les bornes inférieures des questions 3 et 4 se comparent-elles ? Interpréter. *Solution. On a par comparaison immédiate que $\frac{\sum_{i=1}^n t_i^2}{n\sum_{i=1}^n t_i^2 - (\sum_{i=1}^n t_i)^2} \geq \frac{1}{n}$. En effet, on s'attend à bien mieux pouvoir estimer α si l'information sur β est donnée. Cela est compatible avec les bornes obtenues.*

Exercice 4 – Estimation UVMB avec deux observations

On considère un modèle où X_1 et X_2 sont deux variables i.i.d. de même loi, ayant une densité par rapport à la mesure de comptage sur $\mathbb{N}$ donnée par

$$\forall x \in \mathbb{N},\ f_\theta(x) = (x+1)\theta^2(1-\theta)^x,$$

avec θ inconnu dans $\Theta :=]0,1[$.

1. Déterminer $\mathbb{E}_\theta\left[\frac{1}{X_1+1}\right]$ pour tout $\theta \in \Theta$. *Solution.*

$$\mathbb{E}_\theta\left[\frac{1}{X_1+1}\right] = \sum_{x\geq 0}\theta^2(1-\theta)^x = \theta.$$

2. Montrer que $X_1 + X_2$ est une statistique suffisante. *Solution. On écrit la densité jointe de (X_1, X_2). Pour tout $x, y \in \mathbb{N}$, $f_\theta(x,y) = \underbrace{(x+1)(y+1)}_{h(x,y)}\underbrace{\theta^4(1-\theta)^{x+y}}_{=g_\theta(x+y)}$ et on applique le théorème de Neyman-Fisher.*

3. Déterminer la loi de $X_1 + X_2$ et montrer que $X_1 + X_2$ est complète. *Solution. La loi de $X_1 + X_2$ s'obtient par convolution et a pour densité, pour tout $n \in \mathbb{N}$, $g_\theta(n) = \sum_{k=0}^n f_\theta(k)f_\theta(n-k) = \theta^4(1-\theta)^n\sum_{k=0}^n(k+1)(n-k+1) = \frac{1}{6}\theta^4(1-\theta)^n(n+1)(n+2)(n+3)$, après des calculs classiques. La complétude vient ensuite. Soit $g : \mathbb{N} \to \mathbb{R}$ telle que $\mathbb{E}_\theta[g(X_1+X_2)] = 0$ pour tout $\theta \in \Theta$. Cela entraîne, pour tout $\theta \in \Theta$,*

$$\sum_{n\geq 0}(1-\theta)^n(n+1)(n+2)(n+3)g(n) = 0 = \sum_{n\geq 0} 0 \times (1-\theta)^n,$$

par identification des coefficients d'une série entière (en $1-\theta$) sur son disque ouvert de convergence, on a pour tout $n \geq 0$, $(n+1)(n+2)(n+3)g(n) = 0$ donc $g(n) = 0$. La statistique $X_1 + X_2$ est complète.

4. Montrer qu'il existe un unique estimateur UVMB de θ, et donner son expression. *On commencera par déterminer la densité jointe de $(X_1, X_1 + X_2)$.* *Solution. On suit l'indication. Regardons la densité jointe de $(X_1, X_1 + X_2)$. Pour tout $n, m \in \mathbb{N}$, $\mathbb{P}_\theta((X_1, X_1+X_2) = (n, m)) = \mathbb{1}_{m\geq n} f_\theta(n) f_\theta(m-n) = \theta^4(1-\theta)^m(n+1)(m-n+1)\mathbb{1}_{m\geq n}$. Cela donne une densité conditionnelle de X_1 sachant $X_1 + X_2$ qui s'écrit*

$$f_\theta(x \,|\, m) = \mathbb{1}_{x\leq m} \frac{\theta^4(1-\theta)^m(x+1)(m-x+1)}{\frac{1}{6}\theta^4(1-\theta)^m(m+1)(m+2)(m+3)} = \mathbb{1}_{x\leq m} \frac{6(x+1)(m-x+1)}{(m+1)(m+2)(m+3)}\,.$$

On peut dès lors évaluer l'estimateur Rao-Blackwellisé

$$\begin{aligned}
\mathbb{E}_\theta\left[\frac{1}{X_1+1} \,|\, X_1 + X_2 = m\right] &= \sum_{x=0}^{m} \frac{1}{x+1} \frac{6(x+1)(m-x+1)}{(m+1)(m+2)(m+3)} \\
&= \frac{6}{(m+1)(m+2)(m+3)} \sum_{x=0}^{m} (m-x+1) \\
&= \frac{6}{(m+1)(m+2)(m+3)} \frac{(m+1)(m+2)}{2} = \frac{3}{m+3}\,.
\end{aligned}$$

$X_1 + X_2$ étant complète, d'après le cours, il existe un unique estimateur UVMB de θ, donné par

$$\hat{\theta} = \frac{3}{3 + X_1 + X_2}\,.$$

Fin du sujet.