CC2 2024/2025 – Durée 1h30

Les documents et appareils électroniques (calculatrice, téléphone, ordinateur, ...) sont interdits. **Toutes les réponses doivent être justifiées.**

Exercice 1 - Un modèle linéaire

On se donne les observations suivantes : $Y_1 = \mu_1 + \varepsilon_1$, $Y_2 = 2\mu_1 + \varepsilon_2$, $Y_3 = \mu_2 + \varepsilon_3$, $Y_4 = 3\mu_2 + \varepsilon_4$, où les $(\varepsilon_i)_{i=1,\dots,4}$ sont i.i.d. de loi $\mathcal{N}(0, \sigma^2)$, les paramètres μ_1, μ_2 et σ^2 sont inconnus.

On rappelle l'élément d'algèbre linéaire suivant : si $v_1, \dots, v_k$ sont des vecteurs de $\mathbb{R}^d$ (vus comme des vecteurs colonne), et $V = (v_1 | \dots | v_k)$ est la matrice dont les colonnes sont les v_i , alors $\Pi_V = V(V^t V)^{-1} V^t$ est la matrice de la projection orthogonale sur l'espace vectoriel engendré par $v_1, \dots, v_k$ (lorsque $V^t V$ est inversible).

Enfin, on donne les valeurs de quantiles suivantes. On note $\chi^2(p)$ et $\mathcal{T}(p)$ les lois du Chi-deux et de Student à p degrés de liberté.

	2,5%	97,5%
$\chi^2(2)$	0,05	7,38
$\chi^2(3)$	0,22	9,35
$\mathcal{T}(2)$	-4,30	4,30
$\mathcal{T}(3)$	-3,18	3,18

1. Écrire le modèle linéaire correspondant.
2. Donner **explicitement** les estimateurs des moindres carrés de μ_1 et μ_2 , notés $\hat{\mu}_1$ et $\hat{\mu}_2$. Quelles sont leurs lois ?
3. Donner (non explicitement) l'estimateur des moindres carrés de σ^2 , qu'on notera $\hat{\sigma}^2$, quelle est sa loi ?
4. A-t-on indépendance entre $(\hat{\mu}_1, \hat{\mu}_2)$ et $\hat{\sigma}^2$? En déduire des intervalles de niveaux de confiance 95% pour μ_1 et σ^2 .
5. On suppose maintenant que μ_2 est connue. Proposer un estimateur $\tilde{\sigma}^2$ pour σ^2 et donner sa loi.
6. Dans le cas où μ_2 est connue, construire des intervalles de niveaux de confiance 95% basés sur $\tilde{\sigma}^2$ pour σ^2 et μ_1 . Les comparer à ceux obtenus dans le cas où μ_2 est inconnu, sur la base de l'espérance de leur longueur (pour σ^2) et de l'espérance du carré de leur longueur (pour μ_1).

Solution 1 -

1. En notant $Y = (Y_1, \dots, Y_4)^t$ et $\mu = (\mu_1, \mu_2)^t$, on peut écrire le modèle sous la forme

$$Y = X\mu + \varepsilon,$$

avec

$$X = \begin{pmatrix} 1 & 0 \\ 2 & 0 \\ 0 & 1 \\ 0 & 3 \end{pmatrix},$$

et $\varepsilon \sim \mathcal{N}(0_4, \sigma^2 I_4)$.

2. Le cours donne

$$\hat{\mu} = (X^t X)^{-1} X^t Y.$$

Reste à calculer. D'une part

$$X^t X = \begin{pmatrix} 5 & 0 \\ 0 & 10 \end{pmatrix}.$$

Donc

$$(X^t X)^{-1} = \begin{pmatrix} 1/5 & 0 \\ 0 & 1/10 \end{pmatrix}.$$

D'autre part

$$X^t Y = \begin{pmatrix} Y_1 + 2Y_2 \\ Y_3 + 3Y_4 \end{pmatrix}.$$

On en déduit

$$\begin{cases} \hat{\mu}_1 &= \frac{Y_1 + 2Y_2}{5}, \\ \hat{\mu}_2 &= \frac{Y_3 + 3Y_4}{10}. \end{cases}$$

Enfin, d'après le cours (où à la main, ici ça se voit bien),

$$\hat{\mu} \sim \mathcal{N}(0_2, \sigma^2 (X^T X)^{-1}),$$

ce qui ici se traduit par $\hat{\mu}_1 \sim \mathcal{N}(\mu_1, \sigma^2/5)$, $\hat{\mu}_2 \sim \mathcal{N}(\mu_2, \sigma^2/10)$, et $\hat{\mu}_1 \perp\!\!\!\perp \hat{\mu}_2$.

3. En notant $\hat{\varepsilon} = Y - X\hat{\mu} = \pi_{V^\perp} Y$, où V est l'espace engendré par v_1 et v_2 (colonnes de X), le théorème de Cochran indique que

$$\hat{\sigma}^2 = \frac{\|\hat{\varepsilon}\|^2}{2} \sim \sigma^2 \frac{\chi^2(2)}{2}.$$

$\hat{\sigma}^2$ est donc un estimateur non-biaisé de σ^2 (de loi mentionnée au-dessus).

4. En utilisant le Théorème de Cochran là encore, on a que $\pi_V Y \perp\!\!\!\perp \pi_{V^\perp} Y$, et donc $\hat{\sigma}^2 \perp\!\!\!\perp \pi_V(Y)$. Or,

$$\hat{\mu} = (X^t X)^{-1} X^t \pi_V(Y),$$

ce dont on déduit que $\hat{\mu} \perp\!\!\!\perp \hat{\sigma}^2$.

IC sur σ^2 .

On part de

$$\frac{2\hat{\sigma}^2}{\sigma^2} \sim \chi^2(2).$$

En utilisant le tableau de quantiles, on sait qu'avec probabilité 95%, on a

$$0.05 \leq \frac{2\hat{\sigma}^2}{\sigma^2} \leq 7.38,$$

et donc que $\left[\frac{2\hat{\sigma}^2}{7.38}; 40\hat{\sigma}^2\right]$ est un intervalle de niveau de confiance 95%.

IC sur μ_1 .

On part de

$$\frac{\sqrt{5}(\hat{\mu}_1 - \mu_1)}{\hat{\sigma}} \sim \mathcal{T}(2).$$

En utilisant la table de quantiles, $\left[\hat{\mu}_1 - \frac{\hat{\sigma}}{\sqrt{5}}4.3; \hat{\mu}_1 + \frac{\hat{\sigma}}{\sqrt{5}}4.3\right]$ est un intervalle de niveau de confiance 95%.

5. Si μ_2 est connue, on a $\frac{1}{\sigma^2} ((Y_3 - \mu_2)^2 + (Y_4 - 3\mu_2)^2) \sim \chi^2(2)$. En notant $\hat{\varepsilon}_1$ et $\hat{\varepsilon}_2$ les deux premières coordonnées de $\hat{\varepsilon}$, on va remplacer les deux dernières coordonnées par l'expression adéquate contenant μ_2 , c'est à dire

$$\tilde{\sigma}^2 = \frac{1}{3} (\hat{\varepsilon}_1^2 + \hat{\varepsilon}_2^2 + (Y_3 - \mu_2)^2 + (Y_4 - 3\mu_2)^2).$$

Cela revient à remplacer $\hat{\mu}_2$ par μ_2 dans l'expression de σ^2 . Comme $\hat{\varepsilon}_1$ et $\hat{\varepsilon}_2$ sont $\sigma(Y_1, Y_2)$ -mesurables, on a que $\hat{\varepsilon}_1^2 + \hat{\varepsilon}_2^2 \perp\!\!\!\perp (Y_3 - \mu_2)^2 + (Y_4 - 3\mu_2)^2$. Par ailleurs, Cochran encore assure que $\hat{\varepsilon}_1^2 + \hat{\varepsilon}_2^2 \sim \sigma^2 \chi^2(1)$ (considérer le sous-modèle sur $(Y_1, Y_2)^t$ par exemple, où le faire à la main). On en déduit alors que

$$\tilde{\sigma}^2 \sim \sigma^2 \frac{\chi^2(3)}{3}.$$

6. En faisant comme précédemment, on a immédiatement que $\left[\frac{3}{9.35}\tilde{\sigma}^2; \frac{3}{0.22}\tilde{\sigma}^2\right]$ est un intervalle de niveau de confiance 95% sur σ^2 . Sa longueur est inférieure à $15\tilde{\sigma}^2$, là où celle du premier intervalle est supérieure à $39\hat{\sigma}^2$. Comme $\mathbb{E}(\tilde{\sigma}^2) = \mathbb{E}(\hat{\sigma}^2) = \sigma^2$, en espérance ce deuxième intervalle est donc meilleur.

Pour l'IC sur $\hat{\mu}_1$ il faut faire attention aux indépendances. En utilisant Cochran, on a $\hat{\mu}_1 \perp\!\!\!\perp (\hat{\varepsilon}_1, \hat{\varepsilon}_2)$. Comme $(Y_3, Y_4) \perp\!\!\!\perp (\hat{\mu}_1, \hat{\varepsilon}_1, \hat{\varepsilon}_2)$, on a que $\hat{\mu}_1 \perp\!\!\!\perp (\hat{\varepsilon}_1, \hat{\varepsilon}_2, Y_3, Y_4)$, et donc que $\hat{\mu}_1 \perp\!\!\!\perp \tilde{\sigma}^2$. On en déduit alors que

$$\frac{\sqrt{5}(\hat{\mu}_1 - \mu_1)}{\tilde{\sigma}} \sim \mathcal{T}(3),$$

et donc que $\left[\hat{\mu}_1 - \frac{\tilde{\sigma}}{\sqrt{5}}3.18; \hat{\mu}_1 + \frac{\tilde{\sigma}}{\sqrt{5}}3.18; \right]$ est un intervalle de niveau de confiance 95% sur μ_1 . En moyenne, le carré de sa longueur est de $(2 \times 3.18/\sqrt{5})^2 \sigma^2$, là où le carré de la longueur du premier est en moyenne $(2 \times 4.3/\sqrt{5})^2 \sigma^2$. Le second intervalle est donc meilleur là encore (ce qui est assez cohérent avec l'intuition que plus on en sait plus on peut être précis).

Exercice 2 - Comparaison de précisions

Météo France souhaite comparer la précision de deux logiciels de prévision de température, et à cette fin collecte leurs erreurs $(E_{j,i})_{j=1,2,i=1,\dots,n}$ supposées indépendantes, de loi $\mathcal{N}(0, \sigma_j^2)$ (j est le numéro de logiciel, et σ_j est alors la précision du logiciel). On note $\tau = \sigma_1/\sigma_2$ le ratio de ces précisions.

Un statisticien peu futé commence par calculer $X_i = \frac{E_{1,i}}{E_{2,i}}$, et base sa procédure sur l'observation des $X_1, \dots, X_n$.

On rappelle les éléments suivants sur la loi de Cauchy de paramètre a , notée $\mathcal{C}(a)$, où $a > 0$:

— densité (sur $\mathbb{R}$) : $\frac{a}{\pi(a^2+x^2)}$,

— fonction caractéristique : $\phi_{\mathcal{C}(a)}(t) = \exp(-|t|)$.

1. Montrer que si $Y \sim \mathcal{C}(1)$, $aY \sim \mathcal{C}(a)$, et que si $Y_1 \sim \mathcal{C}(a_1)$, $Y_2 \sim \mathcal{C}(a_2)$, avec $Y_1 \perp\!\!\!\perp Y_2$, alors $Y_1 + Y_2 \sim \mathcal{C}(a_1 + a_2)$.
2. On admet que si N_1 et N_2 sont deux lois normales standards indépendantes, $N_1/N_2 \sim \mathcal{C}(1)$. En déduire le modèle pour cette expérience.
3. On note $\hat{\tau}_{n,1} = \bar{X}_n$ (moyenne empirique). $\hat{\tau}_{n,1}$ est-il consistant ?
4. On cherche à prouver que le logiciel 1 est plus précis que le logiciel 2. Donner les hypothèses du test associé.
5. Construire un test $T_{n,1}$ pour ces hypothèses, basé sur $\hat{\tau}_{n,1}$, de niveau α .
6. Ce test est-il plus puissant que $T_{1,1}$ (le même mais basé sur une seule observation) ?
7. Proposez une procédure de test plus futée (et simple), à partir de l'observation des $(E_{j,i})_{j=1,2,i=1,\dots,n}$.
8. Montrer que si N_1 et N_2 sont deux lois normales standards indépendantes, alors $N_1/N_2 \sim \mathcal{C}(1)$.

Solution 2 -

1. Soit $Y \sim \mathcal{C}(1)$, et $t \in \mathbb{R}$. Comme $\phi_{aY}(t) = \phi_Y(at) = \exp(-a|t|) = \phi_{\mathcal{C}(a)}(t)$, on a bien $aY \sim \mathcal{C}(a)$. Soient Y_1 et Y_2 indépendantes et de loi $\mathcal{C}(a_1)$, $\mathcal{C}(a_2)$. Pour tout $t \in \mathbb{R}$ on a

$$\phi_{Y_1+Y_2}(t) = \phi_{Y_1}(t)\phi_{Y_2}(t) = \exp(-(a_1 + a_2)|t|) = \phi_{\mathcal{C}(a_1+a_2)}(t).$$

D'où $Y_1 + Y_2 \sim \mathcal{C}(a_1 + a_2)$.

2. On a $X_1 \sim \frac{\sigma_1 N_1}{\sigma_2 N_2}$, où N_1 et N_2 sont des gaussiennes standards indépendantes. On en déduit alors que $X_1 \sim \tau \mathcal{C}(1) \sim \mathcal{C}(\tau)$, et donc le modèle

$$(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), (\mathcal{C}(\tau)^{\otimes n})_{\tau>0}).$$

3. Comme $\hat{\tau}_{n,1} = \bar{X}_n \sim \frac{1}{n} \mathcal{C}(\tau n) \sim \mathcal{C}(\tau)$, $\hat{\tau}_{n,1}$ ne peut converger en proba (vers quoi que ce soit). En effet, pour tout $x \in \mathbb{R}$ et $\varepsilon > 0$, $P_\tau(|\bar{X}_n - x| \geq \varepsilon) = \mathbb{P}(|\mathcal{C}(1) - (x/\tau)| \geq \varepsilon/\tau) > 0$.
4. On veut prouver $\tau < 1$, on le met donc en H_1 , ce qui donne les hypothèses

$$\begin{cases} H_0 & : \tau \geq 1 \\ H_1 & : \tau < 1. \end{cases}$$

5. Comme $\hat{\tau}_{n,1} \sim \tau \mathcal{C}(1)$, on s'attend à ce que sous H_1 $|\hat{\tau}_{n,1}|$ soit petit. On cherche alors un test de la forme

$$T_{n,1} = \mathbb{1}_{|\hat{\tau}_{n,1}| \leq t_\alpha}.$$

Reste à calibrer le t_α . Soit $\tau \geq 1$, et $t > 0$,

$$\begin{aligned} P_\tau(|\hat{\tau}_{n,1}| \leq t) &= \mathbb{P}(\tau|\mathcal{C}(1)| \leq t) \\ &\leq \mathbb{P}(|\mathcal{C}(1)| \leq t) \\ &= \int_{-t}^t \frac{1}{\pi(1+x^2)} dx = \frac{1}{\pi} (\arctan(t) - \arctan(-t)) = \frac{2}{\pi} \arctan(t). \end{aligned}$$

En posant $t_\alpha = \tan\left(\frac{\pi\alpha}{2}\right)$, on a bien

$$\sup_{\tau \geq 1} P_\tau(T_{n,1} = 1) = \alpha.$$

On remarque au passage que les $|\mathcal{C}(\tau)|$ sont stochastiquement ordonnées (implicitement prouvé ici).

6. On a $T_{1,1} = \mathbb{1}_{|X_1| \geq t_\alpha}$ (le même seuil que dans $T_{n,1}$ car il ne dépend pas de n). Or, $X_1 \sim \bar{X}_n$, donc en termes de loi ces tests sont exactement les mêmes. En termes de puissance, cela se traduit par, pour n'importe quel $\tau < 1$,

$$P_\tau(T_{n,1} = 1) = \mathbb{P}(|\mathcal{C}(1)| \leq t_\alpha/\tau) = P_\tau(T_{1,1} = 1),$$

et donc

$$\inf_{\tau < 1} P_\tau(T_{n,1} = 1) = \inf_{\tau < 1} P_\tau(T_{1,1} = 1).$$

Ces tests sont donc de même puissance (de fait ils ont même loi pour tout τ , ce qui justifie le caractère peu futé attribué au statisticien).

7. Plutôt que faire des sommes de ratios, il est plus malin de faire un ratio de sommes. On pose

$$S = \frac{\sum_{i=1}^n E_{1,i}^2}{\sum_{i=1}^n E_{2,i}^2} \sim \tau^2 \frac{S_1}{S_2},$$

où S_1 et S_2 sont des $\chi^2(n)$ indépendantes. On a alors immédiatement que

$$S \sim \tau^2 \mathcal{F}(n, n),$$

loi de Fischer à (n, n) degrés de libertés. Sous H_1 on s'attend à ce que S soit plus petit que sous H_0 , on cherche donc un test de la forme

$$T_{n,2} = \mathbb{1}_{S \leq t_\alpha}.$$

On calibre facilement le t_α en le prenant égal à q le quantile d'ordre α d'une loi $\mathcal{F}(n, n)$. En effet, pour tout $\tau \geq 1$,

$$P_\tau(S \leq q) = \mathbb{P}(\tau^2 \mathcal{F}(n, n) \leq q) \leq \mathbb{P}(\mathcal{F}(n, n) \leq q) = \alpha,$$

et donc

$$\sup_{\tau \geq 1} P_\tau(T_{n,2} = 1) = \alpha.$$

8. On pose $Z = N_1/N_2$ (bien définie p.s.). Soit f mesurable positive.

$$\mathbb{E}(f(Z)) = \int_{\mathbb{R}^*} \int_{\mathbb{R}^*} \frac{1}{2\pi} f(x/y) e^{-\frac{1}{2}(x^2+y^2)} dx dy.$$

Soit

$$\phi : \begin{cases} \mathbb{R}^* \times \mathbb{R}^* & \rightarrow \mathbb{R}^* \times \mathbb{R}^* \\ (x, y) & \mapsto (x/y, y). \end{cases}$$

ϕ est bien un C^1 -difféomorphisme, de Jacobien $J_\phi(x, y) = |y|^{-1}$. On a alors, en utilisant un changement de variable (avec $u = x/y$, $v = y$),

$$\mathbb{E}(f(Z)) = \int_{\mathbb{R}^*} \int_{\mathbb{R}^*} \frac{f(u)}{2\pi} |v| e^{-\frac{1}{2}(v^2(1+u^2))} du dv.$$

Or, pour tout u ,

$$\int_{\mathbb{R}^*} |v| e^{-\frac{1}{2}(v^2(1+u^2))} dv = 2 \int_0^{+\infty} v e^{-\frac{1}{2}(v^2(1+u^2))} dv = \frac{2}{1+u^2}.$$

On en déduit (en utilisant Fubini positif),

$$\mathbb{E}(f(Z)) = \int_{\mathbb{R}} \frac{f(u)}{\pi(1+u^2)} du,$$

et donc $Z \sim \mathcal{C}(1)$.