CC2 2023/2024 – Durée 1h30

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

Notation. Dans tout le sujet, la moyenne empirique de $z = (z_1, \dots, z_n) \in \mathbb{R}^n$ est notée $\overline{z} := \frac{1}{n}\sum_{i=1}^n z_i$.

Exercice 1 - Estimation par maximum de vraisemblance.

On considère un échantillon $X_1, \dots, X_n$ de variables aléatoires de loi $\mathcal{U}[\theta, \theta+l]$ avec $\theta, l > 0$. On cherche à estimer l.

1. Écrire le modèle considéré. Montrer que son EMV est donné par

$$(\widehat{\theta}, \widehat{l}) = (\min_{i=1,\dots,n} X_i, \max_{i=1,\dots,n} X_i - \min_{i=1,\dots,n} X_i).$$

Correction : Le modèle est $(\mathbb{R}^n, B(\mathbb{R}^N)), (\mathcal{U}[\theta, \theta+l])^n)$. Il est dominé par la mesure de Leb. Sa vraisemblance est $L_n(\theta, l) = l^{-n}\prod_{i=1}^n \mathbf{1}_{[\theta,\theta+l]}(X_i) = l^{-n}\mathbf{1}_{\min\geq\theta, \max\leq\theta+l}$. On a $L_n(\theta, l) \leq (\max - \theta)^{-n}\mathbf{1}_{\min\geq\theta,\max\leq\theta+l} \leq (\max - \min)^{-n} = L_n(\min, \max - \min)$. D'où l'EMV annoncé.

2. Déterminer la loi limite de $n(\min_{i=1,\dots,n} X_i - \theta)$ et de $n(\theta + l - \max_{i=1,\dots,n} X_i)$. En déduire que $\widehat{\theta}$ converge en probabilité vers θ et que $\widehat{l}$ converge en probabilité vers l.

Correction : Pour $t \geq 0$, on trouve que $\mathbb{P}(n(\min_{i=1,\dots,n} X_i - \theta) \geq t) = (\frac{\theta+l-\theta-t/n}{l})^n \to e^{-t/l}$ et $\mathbb{P}(n(l+\theta - \max_{i=1,\dots,n} X_i) \geq t) = (\frac{\theta+l-t/n-\theta}{l})^n \to e^{-t/l}$. Par Slutsky, on a la consistance (ou en prenant $t = \epsilon n$, avant de passer à la limite en n).

3. Montrer que pour tout $(x, y) \in [\theta, \theta+l]^2$:

$$F(x,y) = \mathbb{P}(\min_{i=1,\dots,n} X_i \leq x, \max_{i=1,\dots,n} X_i \leq y) = \left(\frac{y-\theta}{l}\right)^n - \mathbf{1}_{x\leq y}\left(\frac{y-x}{l}\right)^n.$$

En déduire la densité $f(x,y)$ de $(\min_{i=1,\dots,n} X_i, \max_{i=1,\dots,n} X_i)$.

Correction : $F(x,y) = \mathbb{P}(\max \leq y) - \mathbb{P}(\min > x, \max \leq y)$, ce qui donne la formule demandée. Pour la densité, on dérive par rapport à x et y pour obtenir : $f(x,y) = n(n-1)\frac{(y-x)^{n-2}}{l^n}\mathbf{1}_{\theta\leq x\leq y\leq\theta+l}$.

4. Calculer la fonction de répartition de $\widehat{l}/l$. En déduire un intervalle de confiance de niveau $1-\alpha$ pour l (on ne cherchera pas à calculer explicitement les quantiles de $\widehat{l}/l$).

Correction : $\mathbb{P}(\widehat{l} \leq t) = \int_{[\theta,\theta+l]^2}\mathbf{1}_{y-x\leq t}f(x,y)dxdy = \int_\theta^{\theta+l}\int_x^{(x+t)\wedge(\theta+l)} f(x,y)dydx = l^{-n}\int_\theta^{\theta+l-t} nt^{n-1}dx + l^{-n}\int_{\theta+l-t}^{\theta+l} n(\theta+l-x)^{n-1}dx = n(t/l)^{n-1}(1-t/l) + (t/l)^n$. Donc $\widehat{l}/l$ est pivotale. Pour q_r qui vérifie $nq_r^{n-1}(1-q_r) + q_r^n = r$, on a l'IC $[\widehat{l}/q_{1-\alpha/2}, \widehat{l}/q_{\alpha/2}]$.

Exercice 2 - Test de corrélation.

On considère un échantillon $(X_1,Y_1),\dots,(X_n,Y_n)$ où (X_1,Y_1) est gaussien de moyenne $\mu=(\mu_1,\mu_2)^T\in\mathbb{R}^2$ et de matrice de covariance Σ avec :

$$\Sigma := \begin{pmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix},\quad \sigma_1>0,\quad \sigma_2>0,\quad \rho\in(-1,1).$$

On cherche à tester $H_0:\rho=0$ contre $H_1:\rho\neq 0$. Dans toute la suite μ_1,μ_2,σ_1 et σ_2 ne sont pas supposés connus.

1. Pourquoi ce test revient à tester H_0 : X_1 est indépendant de Y_1, contre H_1 : X_1 n'est pas indépendant de Y_1 ?

   Correction : Pour les vecteurs gaussiens décorrélation et indépendances sont équivalents.

2. Donner un estimateur non biaisé $\widehat{\sigma}_1^2$ de σ_1^2 et construire un intervalle de confiance non asymptotique pour σ_1^2 de niveau $1-\alpha\in(0,1)$.

   Correction : On prend $\widehat{\sigma}_1^2=\frac{1}{n-1}\sum_{i=1}^n(X_i-\overline{X})^2=\frac{1}{n-1}\|X-\pi(X)\|^2$ où π est la projection orthogonale sur Vect $(1,\dots,1)$. D'après Cochran, $(n-1)\widehat{\sigma}_1^2$ suit une loi $\sigma_1^2\chi^2(n-1)$. Il est donc non biaisé. L'IC est $[\frac{(n-1)\widehat{\sigma}_1^2}{q_{1-\alpha/2}},\frac{(n-1)\widehat{\sigma}_1^2}{q_{\alpha/2}}]$.

3. Montrer que $\rho=\frac{\mathrm{Cov}(X_1,Y_1)}{\sigma_1\sigma_2}$ puis que $\frac{1}{n}\sum_{i=1}^n(X_i-\overline{X})(Y_i-\overline{Y})=\frac{1}{n}\sum_{i=1}^n X_iY_i-\overline{X}\times\overline{Y}$.

   Correction : Par définition de Σ on a l'expression de ρ. La second expression se démontre en développant.

   Au vu des questions précédentes, on considère donc comme estimateur de ρ :

$$r_n=\frac{\sum_{i=1}^n(X_i-\overline{X})(Y_i-\overline{Y})}{\sqrt{\sum_{i=1}^n(X_i-\overline{X})^2\sum_{i=1}^n(Y_i-\overline{Y})^2}}.$$

4. Montrer que $\rho\in[-1,1]$ et $r_n\in[-1,1]$.

   Correction : On applique 2 fois l'inégalité de Cauchy-Schwarz.

   Pour la suite, on suppose que $r_n\in(-1,1)$.

5. Montrer que r_n converge presque surement vers ρ.

   Correction : On ne change pas r_n en translatant X_1 et Y_1 donc OPS $\mu_1=\mu_2=0$. On décompose : $r_n=\frac{\sum_{i=1}^n X_iY_i}{(n-1)\widehat{\sigma}_1\widehat{\sigma}_2}-\frac{\overline{X}\times\overline{Y}}{\widehat{\sigma}_1\widehat{\sigma}_2}$. Par la LGN, $\overline{X}\times\overline{Y}$ converge ps vers 0 et $\frac{\sum_{i=1}^n X_iY_i}{(n-1)}$ converge ps vers $\rho\sigma_1\sigma_2$. De plus, $\widehat{\sigma}_1{}^2=\frac{1}{n}\sum_{i=1}^n Y_i^2-\overline{Y}^2$ et encore par la LGN ce terme converge vers σ_1^2. De même pour σ_2^2. Par opérations, on a convergence ps demandée.

6. Déterminer la loi limite de $\sqrt{n}r_n$ sous H_0. En déduire un test de niveau $\alpha\in(0,1)$ pour $H_0:\rho=0$ contre $H_1:\rho\neq 0$.

   Correction : $\sqrt{n}r_n=\sqrt{n}\frac{\sigma_1\sigma_2}{\widehat{\sigma}_1\widehat{\sigma}_2}\frac{\sum_{i=1}^n X_iY_i}{(n-1)\sigma_1\sigma_2}-\sqrt{n}\frac{\overline{X}\times\overline{Y}}{\widehat{\sigma}_1\widehat{\sigma}_2}$, Sous H_0, $\mathrm{E}[X_1Y_1]=0$ et la variance est $\sigma_1^2\sigma_2^2$. Le TCL donne que $\sqrt{n}\frac{1}{n-1}\sum_{i=1}^n X_iY_i$ converge en loi vers une

$\mathcal{N}(0, \rho\sigma_1\sigma_2)$. On a vu en 5 que $\frac{\sigma_1\sigma_2}{\widehat{\sigma}_1\widehat{\sigma}_2}$ converge ps donc en proba vers 1. Ainsi, Slustky donnent que le premier terme tend vers une $\mathcal{N}(0,1)$. Le second tend vers 0. En effet, en supposant $\mu_1 = \mu_2 = 0$ comme en 5, on a $\mathrm{Var}(\sqrt{n}\overline{X} \times \overline{Y}) = n\frac{\sigma_1}{n}\frac{\sigma_2}{n}$ qui tend vers 0. Donc $\sqrt{n}\overline{X} \times \overline{Y}$ tend vers 0 en proba. Avec Slutksy on obtient donc la convergence de ce terme vers 0. Finalement, encore avec Slutksy on obtient la convergence en loi de r_n vers une $\mathcal{N}(0,1)$. On prend pour test $T = \mathbf{1}_{|r_n| \geq q/\sqrt{n}}$ avec q le quantile d'ordre $1-\alpha/2$ d'une $\mathcal{N}(0,1)$(on vérifie qu'il est de niveau α).

7. On admet que pour tout $\rho \in (-1,1)$, on a $\sqrt{n}(r_n - \rho) \xrightarrow[n\to\infty]{Loi} \mathcal{N}(0, (1-\rho^2)^2)$. Trouver une fonction $f : (-1,1) \to (-1,1)$ strictement croissante tel que $\sqrt{n}(f(r_n) - f(\rho)) \xrightarrow[n\to\infty]{Loi} \mathcal{N}(0,1)$. En déduire $q > 0$ tel que $f^{-1}([f(r_n) - \frac{q}{\sqrt{n}}, f(r_n) + \frac{q}{\sqrt{n}}])$ soit un intervalle de confiance asymptotique pour ρ de niveau $\alpha \in (0,1)$.

   Correction : On veut $f'(\rho)(1-\rho^2) = 1$. Par décomposition en élément simples, on prend $f(x) = \frac{1}{2}\log(\frac{1+x}{1-x})$, qui est bien strictement croissant par l'EDO. La Δ-methode donne le TCL. L'IC est $f^{-1}([f(\rho) \pm q/\sqrt{n}])$ avec q quantile d'ordre $1-\alpha/2$ d'une normale centrée réduite.

Exercice 3 - Modèle linéaire gaussien.

On considère la régression linéaire : $\forall i = 1, \dots, n, \quad Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ avec $\beta_0, \beta_1 \in \mathbb{R}$, $(x_i)_{1\leq i\leq n} \in \mathbb{R}^n$ et $(\epsilon_i)_{1\leq i\leq n}$ des variables aléatoires i.i.d. normales centrées et de variance s^2. On suppose que $\sum_{i=1}^n (x_i - \overline{x})^2 > 0$. On cherche à tester $H_0 : \beta_1 = 0$ contre $H_1 : \beta_1 \neq 0$.

1. Montrer que les estimateurs des moindres carrés $\widehat{\beta}_0$ et $\widehat{\beta}_1$ sont donnés par $\widehat{\beta}_1 = \frac{\sum_{i=1}^n (Y_i - \overline{Y})(x_i - \overline{x})}{\sum_{i=1}^n (x_i - \overline{x})^2}$ et $\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1\overline{x}$. Quelle est la loi de $\widehat{\beta}_1$?

   Correction : Par minimisation de $f(a,b) = \sum_{i=1}^n (Y_i - ax_i - b)^2$, on obtient : $\widehat{\beta}_1 = \frac{\sum_{i=1}^n (Y_i - \overline{Y})(x_i - \overline{x})}{\sum_{i=1}^n (x_i - \overline{x})^2}$ et $\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1\overline{x}$. $\widehat{\beta}_1$ est gaussien de moyenne β_1 et de variance $s^2/\sum_{i=1}^n (x_i - \overline{x})^2$. En effet, $\mathrm{Var}[\widehat{\beta}_1] = \frac{\sum_{i=1}^n \mathrm{Var}[Y_i](x_i - \overline{x})^2}{\left(\sum_{i=1}^n (x_i - \overline{x})^2\right)^2}$.

2. Soit $\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$. Montrer que $\sum_{i=1}^n (\widehat{Y}_i - \overline{Y})^2$ et $\sum_{i=1}^n (Y_i - \widehat{Y}_i)^2$ sont indépendants.

   Correction : On reconnait le 2nd terme qui est la somme des carré des résidus. Relions le 1er à $(\widehat{\beta}_1)^2$. En utilisant, $\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i = \overline{Y} - \widehat{\beta}_1(\overline{x} - x_i)$, on obtient $\sum_{i=1}^n (\widehat{Y}_i - \overline{Y})^2 = (\widehat{\beta}_1)^2 \sum_{i=1}^n (x_i - \overline{x})^2$. D'après Cochran, $\widehat{\beta}_1$ est indépendant de $\sum_{i=1}^n (Y_i - \widehat{Y}_i)^2$, d'où le résultat.

3. En déduire la loi de $\frac{(n-2)\sum_{i=1}^n (\widehat{Y}_i - \overline{Y})^2}{\sum_{i=1}^n (Y_i - \widehat{Y}_i)^2}$ si $\beta_1 = 0$.

   Correction : D'après les questions précédentes, si $\beta_1 = 0$, le numérateur suit une loi $\chi^2(1)$. D'après Cochran, le dénominateur une loi du $\chi^2(n-2)$. Les deux sont indépendants, donc le tout suit une loi de Fischer $\mathcal{F}(1, n-2)$.

4. En déduire un test de niveau α pour $H_0 : \beta_1 = 0$ contre $H_1 : \beta_1 \neq 0$.

   Correction : On note F la stats de la question précédente. On prend pour test $\mathbf{1}_{T \geq q}$ avec q quantile d'ordre $1-\alpha$ de $\mathcal{F}(1, n-2)$.

Exercice 4 - Test de corrélation non asymptotique.

On utilise les résultats de l'exercice 3, pour construire un test non asymptotique pour l'exercice 2. On considère donc $(X_1, Y_1), \dots, (X_n, Y_n)$ des couples gaussiens de moyenne μ et de covariance Σ (définis dans l'exercice 2).

1. Montrer que $\mathbb{E}[Y_1|X_1] = \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(X_1 - \mu_1)$. Indication : on pourra commencer par trouver $t \in \mathbb{R}$ tel que $X_1 + tY_1$ et X_1 soient indépendants.

   Correction : Si $\rho = 0$ on prend $t = 0$. Sinon, en calculant la covariance, on trouve $t = -\frac{\sigma_1}{\rho\sigma_2}$. Par conséquent $\mu_1 + t\mu_2 = \mathbb{E}[X_1 + tY_1] = \mathbb{E}[X_1 + tY_1|X_1] = X_1 + t\mathbb{E}[Y_1|X_1]$, ce qui donne le résultat.

2. En déduire que $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ avec β_0, β_1 à exprimer en fonction de $\mu_1, \mu_2, \sigma_1, \sigma_2, \rho$ et $(\epsilon_i)_{1 \leq i \leq n}$ i.i.d. de loi $\mathcal{N}(0, s^2)$ où s^2 n'est pas à calculer.

   Correction : On écrit $Y_i = \mathbb{E}[Y_i|X_i] + Y_i - \mathbb{E}[Y_i|X_i] =: \mathbb{E}[Y_i|X_i] + \epsilon_i$ et on identifie avec la 1.

   Pour la suite, on travaille conditionnellement à $X_1 = x_1, \dots, X_n = x_n$.

3. Avec les notations de l'ex 3, mq $\sum_{i=1}^n (Y_i - \overline{Y})^2 = \sum_{i=1}^n (\widehat{Y}_i - \overline{Y})^2 + \sum_{i=1}^n (Y_i - \widehat{Y}_i)^2$.

   Correction : On note π la projection orthogonale sur le le sev engendré par $\mathbf{1} := (1, \dots, 1)^T$ et $(x_1, \dots, x_n)^T$. On note $\mathbf{Y}$ le vecteur de composantes $Y_i - \overline{Y}$. D'après Pythagore : $\|\mathbf{Y} - \overline{Y}\mathbf{1}\|^2 = \|\pi(\mathbf{Y} - \overline{Y}\mathbf{1})\|^2 + \|(id - \pi)(\mathbf{Y} - \overline{Y}\mathbf{1})\|^2 = \|\widehat{\mathbf{Y}} - \overline{Y}\mathbf{1})\|^2 + \|\mathbf{Y} - \widehat{\mathbf{Y}} + 0\|^2$.

4. On note $F = \frac{\sum_{i=1}^n (\widehat{Y}_i - \overline{Y})^2}{\sum_{i=1}^n (Y_i - \widehat{Y}_i)^2}$. Montrer que $F = r_n^2(1 + F)$ où r_n est l'estimateur de ρ défini dans l'exercice 2, pris en $X_1 = x_1, \dots, X_n = x_n$.

   Correction : D'après les calculs de la question 2.ex3. $F = (\widehat{\beta}_1)^2 \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\sum_{i=1}^n (Y_i - \overline{Y})^2} = r_n(x)^2 \frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \overline{Y})^2}$. D'après la question précédente, $F = r_n(x)^2 (1 + \frac{\sum_{i=1}^n (\widehat{Y}_i - \overline{Y})^2}{\sum_{i=1}^n (Y_i - \widehat{Y}_i)^2})$.

5. En déduire que si $\rho = 0$, $\frac{(n-2)r_n}{1 - r_n^2}$ suit conditionnellement à $X_1 = x_1, \dots, X_n = x_n$ une loi de Fischer $\mathcal{F}(1, n-2)$. Pourquoi $\frac{(n-2)r_n^2}{1 - r_n^2}$ suit alors une loi de Fischer $\mathcal{F}(1, n-2)$? Donnez un test non asymptotique de niveau α pour $H_0 : \rho = 0$ contre $H_1 : \rho \neq 0$.

   Correction : D'après la question précédente $\frac{(n-2)r_n}{1 - r_n^2} = \frac{(n-2)) \sum_{i=1}^n (\widehat{Y}_i - \overline{Y})^2}{\sum_{i=1}^n (Y_i - \widehat{Y}_i)^2}$. On a vu dans l'exercice 3 que cette dernière quantité suit une loi de Fischer $\mathcal{F}(1, n-2)$ si $\beta_1 = 0$,conditionnellement à $X_1 = x_1, \dots, X_n = x_n$. Or $\beta_1 = 0$ ssi $\rho = 0$. Cette loi ne dépend pas de $x_1, \dots, x_n$, donc la conclusion précédente reste vraie sans conditionnement. Le test non asymptotique est le même qu'en 4.ex3.

6. (Bonus) Que proposez-vous pour démontrer le résultat admis en question 7 de l'exercice 2 : $\sqrt{n}(r_n - \rho) \xrightarrow[n \to \infty]{Loi} \mathcal{N}(0, (1 - \rho^2)^2)$?

   Correction : On fait une Δ-methode $\sqrt{n}(g(\frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y}), \widehat{\sigma}_1, \widehat{\sigma}_2) - g(\rho\sigma_1\sigma_2, \sigma_1, \sigma_2))$ avec $g(x, y, z) = \frac{x}{\sqrt{yz}}$.