

Troisième Année Licence M.I.A.S.H.S. 2023 – 2024

Statistique 2

Contrôle Continu 2, Avril 2024

Examen de 1h30. Tout document ou calculatrice est interdit.

Exercice 1 (Sur 12 points)

 Dans la suite, pour un vecteur $U = (U_1, \dots, U_n) \in \mathbf{R}^n$ on note $\|U\|^2 = \sum_{k=1}^n U_k^2$.

- Soit $Z_1 \stackrel{\mathcal{L}}{\sim} \mathcal{N}(0, 1)$ et soit g une fonction de classe \mathcal{C}^1 sur \mathbf{R} et $C > 0$ tel que $|xg(x)| + |g'(x)| \leq C(1+x^2)^{-1}e^{x^2/2}$ pour tout $x \in \mathbf{R}$. Montrer que $\mathbb{E}[|g'(Z_1)| + |Z_1g(Z_1)|] < \infty$ (**1pt**), puis que $\mathbb{E}[g'(Z_1)] = \mathbb{E}[Z_1g(Z_1)]$ (**1.5pts**). En déduire que $\mathbb{E}[Z_1^4] = 3$ (**1pt**).
- Soit $X_1 \stackrel{\mathcal{L}}{\sim} \mathcal{N}(\mu_1, \sigma^2)$ avec $\mu_1 \in \mathbf{R}$ et $\sigma^2 > 0$. Démontrer que $\mathbb{E}[g'(X_1)] = \frac{1}{\sigma^2} \mathbb{E}[(X_1 - \mu_1)g(X_1)]$ (**1.5pts**).
- Soit $(Z_k)_{k \in \mathbf{N}^*}$ une suite de v.a.i.i.d. de même loi $\mathcal{N}(0, 1)$. Déterminer la loi de $Z = (Z_1, \dots, Z_n)$ où $n \geq 1$ (**0.5pts**). Si $h : \mathbf{R}^n \rightarrow \mathbf{R}$ est une fonction de classe \mathcal{C}^1 sur \mathbf{R}^n telle que $\mathbb{E}[\|\nabla h(Z)\| + \|Zh(Z)\|] < \infty$ avec $\nabla h(x) = \left(\frac{\partial}{\partial x_i} h(x)\right)_{1 \leq i \leq n}$, démontrer que $\mathbb{E}[\nabla h(Z)] = \mathbb{E}[Zh(Z)]$ (**3pts**).
- Soit $(X_k)_{k \in \mathbf{N}^*}$ une suite de v.a. indépendantes, telle que $X_k \stackrel{\mathcal{L}}{\sim} \mathcal{N}(\mu_k, \sigma^2)$ avec $\mu_k \in \mathbf{R}$ pour tout $k \in \mathbf{N}^*$. Déterminer la loi de $X = (X_1, \dots, X_n)$ où $n \geq 1$ (**0.5pts**). Soit $h : \mathbf{R}^n \rightarrow \mathbf{R}$ une fonction de classe \mathcal{C}^1 sur \mathbf{R}^n telle que $\mathbb{E}[\|\nabla h(X)\| + \|Xh(X)\|] < \infty$. Démontrer que $\sigma^2 \mathbb{E}[\nabla h(X)] = \mathbb{E}[(X - \mu)h(X)]$ en spécifiant μ (**1.5pts**). En déduire que si $h : (x_1, \dots, x_n) \in \mathbf{R}^n \mapsto (h_1(x_1, \dots, x_n), \dots, h_n(x_1, \dots, x_n))$ une fonction de classe \mathcal{C}^1 sur \mathbf{R}^n telle que $\mathbb{E}[\|\nabla h_i(X)\| + \|X_i h_i(X)\|] < \infty$ pour tout i , alors

$$\frac{1}{\sigma^2} \sum_{i=1}^n \text{cov}(X_i, h_i(X)) = \sum_{i=1}^n \mathbb{E}\left[\frac{\partial}{\partial x_i} h_i(X)\right] \quad (\mathbf{1.5pts}). \quad (1)$$

Proof. 1. On a $\mathbb{E}[|g'(Z_1)|] = \frac{1}{\sqrt{2\pi}} \int_{\mathbf{R}} |g'(x)| e^{-x^2/2} dx \leq \frac{C}{\sqrt{2\pi}} \int_{\mathbf{R}} (1+x^2)^{-1} dx = \frac{C\pi}{\sqrt{2\pi}} < \infty$. Même chose pour $\mathbb{E}[|Z_1g(Z_1)|]$.

On a $\mathbb{E}[g'(Z_1)] = \frac{1}{\sqrt{2\pi}} \int_{\mathbf{R}} g'(x) e^{-x^2/2} dx$. Et avec une intégration par parties

$$\mathbb{E}[Z_1g(Z_1)] = \frac{1}{\sqrt{2\pi}} \int_{\mathbf{R}} g(x) x e^{-x^2/2} dx = \frac{1}{\sqrt{2\pi}} \left(\left[-g(x) e^{-x^2/2} \right]_{-\infty}^{\infty} + \int_{\mathbf{R}} g'(x) e^{-x^2/2} dx \right).$$

Comme $|xg(x)| \leq C(1+x^2)^{-1}e^{x^2/2}$, alors $g(x)e^{-x^2/2} \rightarrow 0$ quand $x \rightarrow \pm\infty$, d'où le résultat.

Considérons la fonction $g(x) = x^3$. Alors on a $g'(x) = 3x^2$ et on a bien $|xg(x)| + |g'(x)| \leq C(1+x^2)^{-1}e^{x^2/2}$ pour tout $x \in \mathbf{R}$. Alors on en déduit que $\mathbb{E}[Z_1^4] = \mathbb{E}[Z_1g(Z_1)] = 3\mathbb{E}[Z_1^2]$. Mais $\mathbb{E}[Z_1^2] = 1$, donc $\mathbb{E}[Z_1^4] = 3$.

- On a $\frac{1}{\sigma^2} \mathbb{E}[(X_1 - \mu_1)g(X_1)] = \frac{1}{\sqrt{2\pi}\sigma} \int_{\mathbf{R}} \frac{(x - \mu_1)}{\sigma} g(x) e^{-(x - \mu_1)^2/2\sigma^2} dx = \frac{1}{\sqrt{2\pi}} \int_{\mathbf{R}} z g(\mu_1 + \sigma z) e^{-z^2/2} dz$ par changement de variable. D'où par intégration par parties,

$$\frac{1}{\sigma^2} \mathbb{E}[(X_1 - \mu_1)g(X_1)] = \frac{1}{\sqrt{2\pi}} \left(\left[g(\mu_1 + \sigma z) e^{-z^2/2} \right]_{-\infty}^{\infty} + \sigma \int_{\mathbf{R}} g'(\mu_1 + \sigma z) e^{-z^2/2} dz \right) = \frac{\sigma}{\sqrt{2\pi}} \int_{\mathbf{R}} g'(\mu_1 + \sigma z) e^{-z^2/2} dz,$$

avec le terme entre crochet qui est égale à 0 comme différence de 2 termes tendant vers 0 en $\pm\infty$. Il ne reste plus qu'à effectuer un changement de variable $x = \mu_1 + \sigma z$ et on a :

$$\frac{\sigma}{\sqrt{2\pi}} \int_{\mathbf{R}} g'(\mu_1 + \sigma z) e^{-z^2/2} dz = \frac{1}{\sqrt{2\pi}} \int_{\mathbf{R}} g'(x) e^{-(x - \mu_1)^2/2\sigma^2} dx = \mathbb{E}[g'(X_1)].$$

- Il est clair que comme $(Z_k)_{k \in \mathbf{N}^*}$ une suite de v.a.i.i.d. de même loi $\mathcal{N}(0, 1)$ alors $Z = (Z_1, \dots, Z_n)$ est un vecteur gaussien de loi $\mathcal{N}_n(0, I_n)$, où I_n est la matrice identité de taille n .

On a

$$\mathbb{E}[Zh(Z)] = \left(\frac{1}{(2\pi)^{n/2}} \int_{\mathbf{R}^n} z_j h(z_1, \dots, z_n) \exp\left(-\frac{1}{2} \sum_{k=1}^n z_k^2\right) dz_1 \dots dz_n \right)_{1 \leq j \leq n}.$$

En utilisant Fubini, on peut d'abord intégrer par rapport à z_j et on obtient par intégration par parties:

$$\begin{aligned} (\mathbb{E}[Z h(Z)])_j &= \frac{1}{(2\pi)^{n/2}} \int_{\mathbf{R}^{n-1}} \exp\left(-\frac{1}{2} \sum_{k=1, k \neq j}^n z_k^2\right) \left(\int_{\mathbf{R}} z_j h(z_1, \dots, z_n) \exp\left(-\frac{1}{2} z_j^2\right) dz_j \right) dz_1 \dots dz_{j-1} dz_{j+1} dz_n \\ &= \frac{1}{(2\pi)^{n/2}} \int_{\mathbf{R}^{n-1}} \exp\left(-\frac{1}{2} \sum_{k=1, k \neq j}^n z_k^2\right) \left(\int_{\mathbf{R}} \frac{\partial}{\partial z_j} h(z_1, \dots, z_n) \exp\left(-\frac{1}{2} z_j^2\right) dz_j \right) dz_1 \dots dz_{j-1} dz_{j+1} dz_n \\ &= \frac{1}{(2\pi)^{n/2}} \int_{\mathbf{R}^n} \frac{\partial}{\partial z_j} h(z_1, \dots, z_n) \exp\left(-\frac{1}{2} \sum_{k=1}^n z_k^2\right) dz_1 \dots dz_n = \mathbb{E}\left[\frac{\partial}{\partial z_j} h(Z)\right], \end{aligned}$$

d'où le résultat.

4. Grâce à l'indépendance, $X = (X_1, \dots, X_n)$ est un vecteur gaussien et $X \stackrel{\mathcal{L}}{\sim} \mathcal{N}_n(\mu, \sigma^2 I_n)$, où $\mu = (\mu_1, \dots, \mu_n)$.

Même différence de preuve qu'au 2., avec changement de variable dans l'intégrale.

On considère donc maintenant une fonction $h = (h_1, \dots, h_n)$. D'après la propriété précédente, $\sigma^2 \mathbb{E}\left[\frac{\partial}{\partial x_i} h_i(X)\right] = \mathbb{E}[(X_i - \mu_i) h_i(X)]$

et comme $\mathbb{E}[X_i] = \mu_i$, on en déduit que $\mathbb{E}[(X_i - \mu_i) h_i(X)] = \text{cov}(X_i, h_i(X))$. D'où le résultat en sommant. \square

Exercice 2 (Sur 13 points)

On considère $(X_k)_{k \in \mathbf{N}^*}$ une suite de v.a. indépendantes, telle que $X_k \stackrel{\mathcal{L}}{\sim} \mathcal{N}(m_k, 1)$ avec $m_k \in \mathbf{R}$ pour tout $k \in \mathbf{N}^*$.

On suppose que (X_1, \dots, X_n) a été observé et $m = (m_1, \dots, m_n)$ est un vecteur inconnu (avec $n \geq 2$).

1. Déterminer la vraisemblance du modèle statistique, après avoir précisé ce dernier **(1pt)**.
2. Prouver qu'il existe un unique estimateur \hat{m} de m par maximum de vraisemblance et $\hat{m} = (X_1, \dots, X_n)$ **(1pt)**.
3. Déterminer le biais de \hat{m} **(0.5pts)** et montrer que son risque quadratique est $R(\hat{m}) = n$ **(1pt)**.
4. Soit maintenant l'estimateur $\tilde{m} = (\tilde{m}_1, \dots, \tilde{m}_n)$ tel que

$$\tilde{m}_k = \tilde{m}_k(X_1, \dots, X_n) = X_k \left(1 - \frac{n-2}{\sum_{i=1}^n X_i^2}\right) \text{ pour tout } k \in \{1, \dots, n\}.$$

Déterminer la limite en loi de \tilde{m}_k quand $n \rightarrow \infty$ **(2pts)**.

5. Démontrer que $R(\tilde{m}) = -n + \mathbb{E}[\|\tilde{m} - X\|^2] + 2 \sum_{i=1}^n \text{cov}(X_i, \tilde{m}_i)$ **(2.5pts)**. En utilisant (1), en déduire que

$$R(\tilde{m}) = -n + \mathbb{E}[\|\tilde{m} - X\|^2] + 2 \sum_{i=1}^n \mathbb{E}\left[\frac{\partial}{\partial x_i} \tilde{m}_i(X)\right] \text{ (1pt)}.$$

6. Montrer que $\mathbb{E}[\|\tilde{m} - X\|^2] = \mathbb{E}\left[\frac{(n-2)^2}{\sum_{i=1}^n X_i^2}\right]$ **(1pt)** et $\sum_{i=1}^n \frac{\partial}{\partial x_i} \tilde{m}_i(X) = n - \frac{(n-2)^2}{\sum_{i=1}^n X_i^2}$ **(1.5pts)**. En déduire que

$$R(\tilde{m}) = n - (n-2)^2 \mathbb{E}\left[\frac{1}{\sum_{i=1}^n X_i^2}\right] \text{ (1pt)}.$$

Comment expliquer que $R(\tilde{m}) < R(\hat{m})$ pour $n \geq 3$ **(0.5pts)**?

Proof. 1. Le modèle statistique est $(\mathbf{R}^n, \mathcal{B}(\mathbf{R}^n), \mathcal{N}_n(m, I_n), m \in \mathbf{R}^n)$.

Le modèle est dominé par la mesure de Lebesgue sur \mathbf{R}^n et sa vraisemblance est pour $(x_1, \dots, x_n) \in \mathbf{R}^n$:

$$V_m(x_1, \dots, x_n) = \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2} \sum_{k=1}^n (x_k - m_k)^2\right).$$

2. On peut maximiser la log-vraisemblance en (X_1, \dots, X_n) ce qui revient à maximiser:

$$\hat{L}(m) = -\frac{n}{2} \log(2\pi) - \frac{1}{2} \sum_{k=1}^n (X_k - m_k)^2.$$

La fonction est séparable et son maximum est atteint quand $\sum_{k=1}^n (X_k - m_k)^2$ est minimum, donc clairement pour $\hat{m}_k = X_k$.

3. On a $\mathbb{E}[\hat{m}] = \mathbb{E}[X] = m$: l'estimateur est sans biais.

Le risque quadratique de \hat{m} est $R(m) = \mathbb{E}[\|\hat{m} - m\|^2] = \sum_{k=1}^n \text{var}(X_k) = n$.

4. La suite des (X_k^2) étant une suite de v.a.i.i.d. telle que $\mathbb{E}[X_1^2] = 1 < \infty$, la loi forte des grands nombres permet d'écrire que $\frac{1}{n} \sum_{i=1}^n X_i^2 \xrightarrow[n \rightarrow \infty]{p.s.} \mathbb{E}[X_1^2] = 1$. Par le Lemme de Slutsky, comme $n/(n-2) \xrightarrow[n \rightarrow \infty]{} 1$, donc on a également $\frac{1}{n-2} \sum_{i=1}^n X_i^2 \xrightarrow[n \rightarrow \infty]{p.s.} 1$. La fonction $h(x) = 1/x$ étant continue au voisinage de 1, on en déduit que $\frac{n-2}{\sum_{i=1}^n X_i^2} \xrightarrow[n \rightarrow \infty]{p.s.} 1$. Par conséquent $\tilde{m}_k \xrightarrow[n \rightarrow \infty]{p.s.} 0$ pour tout $k = 1, \dots, n$.

5. On a $R(\tilde{m}) = \mathbb{E}[\|\tilde{m} - m\|^2] = \mathbb{E}[\|(\tilde{m} - \hat{m}) + (\hat{m} - m)\|^2] = \mathbb{E}[\|\tilde{m} - X\|^2] + n + 2\mathbb{E}[(\tilde{m} - \hat{m})(\hat{m} - m)]$. Mais on peut écrire que: $\mathbb{E}[(\tilde{m} - \hat{m})(\hat{m} - m)] = \mathbb{E}[(\tilde{m} - m)(\hat{m} - m)] - \mathbb{E}[(\hat{m} - m)(\hat{m} - m)] = \sum_{i=1}^n \text{cov}(X_i, \tilde{m}_i) - n$, d'où le résultat.

Considérons la fonction $h(X_1, \dots, X_n) = (\tilde{m}_1(X_1, \dots, X_n), \dots, \tilde{m}_n(X_1, \dots, X_n))$. Cette fonction est de classe \mathcal{C}^1 en dehors de $(0, \dots, 0)$. On peut donc appliquer l'égalité (1) avec $\sigma^2 = 1$ et on obtient le résultat demandé.

6. On a $\mathbb{E}[\|\tilde{m} - X\|^2] = (n-2)^2 \mathbb{E}\left[\frac{\sum_{k=1}^n X_k^2}{(\sum_{i=1}^n X_i^2)^2}\right] = (n-2)^2 \mathbb{E}\left[\frac{1}{\sum_{i=1}^n X_i^2}\right]$.

Par dérivation, $\frac{\partial}{\partial x_i} \tilde{m}_i(X) = 1 - (n-2) \frac{1}{\sum_{k=1}^n X_k^2} + 2(n-2) \frac{X_i^2}{(\sum_{k=1}^n X_k^2)^2}$. On en déduit donc que:

$$\sum_{i=1}^n \frac{\partial}{\partial x_i} \tilde{m}_i(X) = n - n(n-2) \frac{1}{\sum_{k=1}^n X_k^2} + 2(n-2) \frac{\sum_{i=1}^n X_i^2}{(\sum_{k=1}^n X_k^2)^2} = n - (n-2)^2 \frac{1}{\sum_{i=1}^n X_i^2}.$$

En rassemblant les morceaux, on obtient:

$$\begin{aligned} R(\tilde{m}) &= -n + \mathbb{E}[\|\tilde{m} - X\|^2] + 2 \sum_{i=1}^n \mathbb{E}\left[\frac{\partial}{\partial x_i} \tilde{m}_i(X)\right] \\ &= -n + (n-2)^2 \mathbb{E}\left[\frac{1}{\sum_{i=1}^n X_i^2}\right] + 2 \mathbb{E}\left[n - (n-2)^2 \frac{1}{\sum_{i=1}^n X_i^2}\right] \\ &= n - (n-2)^2 \mathbb{E}\left[\frac{1}{\sum_{i=1}^n X_i^2}\right]. \end{aligned}$$

□