**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 $\|U\|^2 = \sum_{k=1}^n U_i^2$.

1. Soit $Z_1 \overset{\mathcal{L}}{\sim} \mathcal{N}(0,1)$ et soit $g$ une fonction de classe $\mathcal{C}^1$ sur $\mathbf{R}$ et $C > 0$ tel que $|x\, g(x)| + |g'(x)| \leq C\,(1+x^2)^{-1}\, e^{x^2/2}$ pour tout $x \in \mathbf{R}$. Montrer que $\mathbb{E}[|g'(Z_1)| + |Z_1\, g(Z_1)|] < \infty$ **(1pt)**, puis que $\mathbb{E}[g'(Z_1)] = \mathbb{E}[Z_1\, g(Z_1)]$ **(1.5pts)**. En déduire que $\mathbb{E}[Z_1^4] = 3$ **(1pt)**.

2. Soit $X_1 \overset{\mathcal{L}}{\sim} \mathcal{N}(\mu_1, \sigma^2)$ avec $\mu_1 \in \mathbf{R}$ et $\sigma^2 > 0$. Démontrer que $\mathbb{E}[g'(X_1)] = \dfrac{1}{\sigma^2}\, \mathbb{E}[(X_1 - \mu_1)\, g(X_1)]$ **(1.5pts)**.

3. 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 $h : \mathbf{R}^n \to \mathbf{R}$ est une fonction de classe $\mathcal{C}^1$ sur $\mathbf{R}^n$ telle que $\mathbb{E}[\|\nabla h(Z)\| + \|Z\, h(Z)\|] < \infty$ avec $\nabla h(x) = \Big(\dfrac{\partial}{\partial x_i} h(x)\Big)_{1\leq i\leq n}$, démontrer que $\mathbb{E}[\nabla h(Z)] = \mathbb{E}[Z\, h(Z)]$ **(3pts)**.

4. Soit $(X_k)_{k\in\mathbf{N}^*}$ une suite de v.a. indépendantes, telle que $X_k \overset{\mathcal{L}}{\sim} \mathcal{N}(\mu_k, \sigma^2)$ 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 $h : \mathbf{R}^n \to \mathbf{R}$ une fonction de classe $\mathcal{C}^1$ sur $\mathbf{R}^n$ telle que $\mathbb{E}[\|\nabla h(X)\| + \|X\, h(X)\|] < \infty$. Démontrer que $\sigma^2\, \mathbb{E}[\nabla h(X)] = \mathbb{E}[(X - \mu)\, h(X)]$ en spécifiant $\mu$ **(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

$$\frac{1}{\sigma^2} \sum_{i=1}^n \mathrm{cov}(X_i, h_i(X)) = \sum_{i=1}^n \mathbb{E}\Big[\frac{\partial}{\partial x_i} h_i(X)\Big] \quad \textbf{(1.5pts)}. \tag{1}$$

*Proof.* 1. On a $\mathbb{E}[|g'(Z_1)|] = \dfrac{1}{\sqrt{2\,\pi}} \displaystyle\int_{\mathbf{R}} |g'(x)|\, e^{-x^2/2} dx \leq \frac{C}{\sqrt{2\,\pi}} \int_{\mathbf{R}} (1+x^2)^{-1} dx = \frac{C\,\pi}{\sqrt{2\,\pi}} < \infty$. Même chose pour $\mathbb{E}[|Z_1\, g(Z_1)|]$.
On a $\mathbb{E}[g'(Z_1)] = \dfrac{1}{\sqrt{2\,\pi}} \displaystyle\int_{\mathbf{R}} g'(x)\, e^{-x^2/2} dx$. Et avec une intégration par parties

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

Comme $|x\, g(x)| \leq C\,(1+x^2)^{-1}\, e^{x^2/2}$, alors $g(x)\, e^{-x^2/2} \to 0$ quand $x \to \pm\infty$, d'où le résultat.
Considérons la fonction $g(x) = x^3$. Alors on a $g'(x) = 3x^2$ et on a bien $|x\, g(x)| + |g'(x)| \leq C\,(1+x^2)^{-1}\, e^{x^2/2}$ pour tout $x \in \mathbf{R}$. Alors on en déduit que $\mathbb{E}[Z_1^4] = \mathbb{E}[Z_1 g(Z_1)] = 3\,\mathbb{E}[Z_1^2]$. Mais $\mathbb{E}[Z_1^2] = 1$, donc $\mathbb{E}[Z_1^4] = 3$.

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

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

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:

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

3. 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

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

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

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

d'où le résultat.

4. Grâce à l'indépendance, $X=(X_1,\dots,X_n)$ est un vecteur gaussien et $X \overset{\mathcal{L}}{\sim} \mathcal{N}_n\big(\mu\,,\,\sigma^2\,I_n\big)$, 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, $\sigma^2\,\mathbb{E}\big[\frac{\partial}{\partial x_i}h_i(X)\big]=\mathbb{E}\big[(X_i-\mu_i)\,h_i(X)\big]$ et comme $\mathbb{E}[X_i]=\mu_i$, on en déduit que $\mathbb{E}\big[(X_i-\mu_i)\,h_i(X)\big]=\mathrm{cov}(X_i,h_i(X))$. D'où le résultat en sommant.

□

**Exercice 2 (Sur 13 points)**

On considère $(X_k)_{k\in\mathbf{N}^*}$ une suite de v.a. indépendantes, telle que $X_k \overset{\mathcal{L}}{\sim} \mathcal{N}(m_k,1)$ 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 $\widehat{m}$ de $m$ par maximum de vraisemblance et $\widehat{m}=(X_1,\dots,X_n)$ **(1pt)**.
3. Déterminer le biais de $\widehat{m}$ **(0.5pts)** et montrer que son risque quadratique est $R(\widehat{m})=n$ **(1pt)**.
4. Soit maintenant l'estimateur $\widetilde{m}=(\widetilde{m}_1,\dots,\widetilde{m}_n)$ tel que

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

On suppose $(m_k)_{k\in\mathbf{N}}$ telle que $\dfrac{1}{n}\displaystyle\sum_{k=1}^{n}m_k^2 \underset{n\to\infty}{\longrightarrow} M$. Déterminer la limite en loi de $\widetilde{m}_k$ quand $n\to\infty$ **(2pts)**.

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

$$R(\widetilde{m})=-n+\mathbb{E}[\|\widetilde{m}-X\|^2]+2\sum_{i=1}^{n}\mathbb{E}\Big[\frac{\partial}{\partial x_i}\widetilde{m}_i(X)\Big]\quad\textbf{(1pt)}.$$

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

$$R(\widetilde{m})=n-(n-2)^2\mathbb{E}\Big[\frac{1}{\sum_{i=1}^n X_i^2}\Big]\quad\textbf{(1pt)}.$$

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

*Proof.* 1. Le modèle statistique est $\big(\mathbf{R}^n,\mathcal{B}(\mathbf{R}^n),\mathcal{N}_n\big(m\,,\,I_n\big),m\in\mathbf{R}^n\big)$.
Le modèle est dominé par la mesure de Lebesgue sur $\mathbf{R}^n$ et sa vraisemblance est pour $(x_1,\dots,x_n)\in\mathbf{R}^n$:

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

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

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

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

3. On a $\mathbb{E}[\widehat{m}]=\mathbb{E}[X]=m$: l'estimateur est sans biais.
Le risque quadratique de $\widehat{m}$ est $R(m)=\mathbb{E}\big[\big\|\widehat{m}-m\big\|^2\big]=\sum_{k=1}^n\mathrm{var}(X_k)=n$.

4. La suite des $(X_k^2)$ est une suite de v.a.i. telle que $\mathbb{E}[X_k^2] = m_k^2 + 1$ pour tout $k \in \mathbf{N}$. Zoit $Z_k = X_k - m_k$, donc $Z_k \overset{\mathcal{L}}{\sim} \mathcal{N}(0,1)$ et $(Z_k)$ et $(Z_k^2)$ sont des suites de v.a.i.i.d. d'espérance finie. Ainsi:

$$\begin{aligned}\mathbb{E}\Big[\frac{1}{n}\sum_{i=1}^n X_i^2\Big] &= \frac{1}{n}\sum_{i=1}^n \mathbb{E}\big[m_i^2 + 2\,m_i\,Z_i + Z_i^2\big] = 1 + \frac{1}{n}\sum_{i=1}^n m_i^2\\ \mathrm{var}\Big(\frac{1}{n}\sum_{i=1}^n X_i^2\Big) &= \frac{1}{n^2}\sum_{i=1}^n \mathrm{var}\big(m_i^2 + 2\,m_i\,Z_i + Z_i^2\big) = \frac{2}{n} + \frac{4}{n^2}\sum_{i=1}^n m_i^2 \underset{n\to\infty}{\longrightarrow} 0,\end{aligned}$$

car $\mathrm{cov}(Z_i, Z_i^2) = 0$. Par l'inégalité de Bienaymé-Tchebychev, pour tout $\varepsilon > 0$,

$$\mathbb{P}\Big(\Big|\frac{1}{n}\sum_{i=1}^n X_i^2 - \Big(1 + \frac{1}{n}\sum_{i=1}^n m_i^2\Big)\Big| \geq \varepsilon\Big) \leq \frac{1}{\varepsilon^2}\,\mathrm{var}\Big(\frac{1}{n}\sum_{i=1}^n X_i^2\Big) \underset{n\to\infty}{\longrightarrow} 0.$$

On en déduit donc (Lemme de Slutsky) que:

$$\frac{1}{n}\sum_{i=1}^n X_i^2 - \Big(1 + \frac{1}{n}\sum_{i=1}^n m_i^2\Big) \underset{n\to+\infty}{\overset{\mathcal{P}}{\longrightarrow}} 0 \quad\Longrightarrow\quad \frac{1}{n}\sum_{i=1}^n X_i^2 \underset{n\to+\infty}{\overset{\mathcal{P}}{\longrightarrow}} 1 + M.$$

Par le Lemme de Slutsky, comme $n/(n-2) \underset{n\to\infty}{\longrightarrow} 1$, donc on a également $\dfrac{1}{n-2}\displaystyle\sum_{i=1}^n X_i^2 \underset{n\to+\infty}{\overset{\mathcal{P}}{\longrightarrow}} 1 + M$. La fonction $h(x) = 1 - 1/x$ étant continue au voisinage de $1+M$, on en déduit que $1 - \dfrac{n-2}{\sum_{i=1}^n X_i^2} \underset{n\to+\infty}{\overset{\mathcal{P}}{\longrightarrow}} \dfrac{M}{1+M}$. Par conséquent, toujours en utilisant le Lemme de Slutsky, $\widetilde{m}_k \underset{n\to\infty}{\overset{\mathcal{L}}{\longrightarrow}} \mathcal{N}\Big(\dfrac{M}{1+M}\,m_k\,,\,\dfrac{M^2}{(1+M)^2}\Big)$ pour tout $k = 1, \ldots, n$.

5. On a $R(\widetilde{m}) = \mathbb{E}\big[\big\|\widetilde{m} - m\big\|^2\big] = \mathbb{E}\big[\big\|(\widetilde{m} - \widehat{m}) + (\widehat{m} - m)\big\|^2\big] = \mathbb{E}\big[\|\widetilde{m} - X\|^2\big] + n + 2\,\mathbb{E}\big[{}^t(\widetilde{m} - \widehat{m})\,(\widehat{m} - m)\big]$. Mais on peut écrire que: $\mathbb{E}\big[{}^t(\widetilde{m} - \widehat{m})\,(\widehat{m} - m)\big] = \mathbb{E}\big[{}^t(\widetilde{m} - m)\,(\widehat{m} - m)\big] - \mathbb{E}\big[{}^t(\widehat{m} - m)\,(\widehat{m} - m)\big] = \sum_{i=1}^n \mathrm{cov}(X_i, \widetilde{m}_i) - n$, d'où le résultat.
Considérons la fonction $h(X_1, \ldots, X_n) = \big(\widetilde{m}_1(X_1, \ldots, X_n), \ldots, \widetilde{m}_n(X_1, \ldots, X_n)\big)$. Cette fonction est de classe $\mathcal{C}^1$ en dehors de $(0, \ldots, 0)$. On peut donc appliquer l'égalité (1) avec $\sigma^2 = 1$ et on obtient le résultat demandé.

6. On a $\mathbb{E}\big[\|\widetilde{m} - X\|^2\big] = (n-2)^2\mathbb{E}\Big[\dfrac{\sum_{k=1}^n X_k^2}{\big(\sum_{i=1}^n X_i^2\big)^2}\Big] = (n-2)^2\mathbb{E}\Big[\dfrac{1}{\sum_{i=1}^n X_i^2}\Big]$.
Par dérivation, $\dfrac{\partial}{\partial x_i}\widetilde{m}_i(X) = 1 - (n-2)\dfrac{1}{\sum_{k=1}^n X_k^2} + 2(n-2)\dfrac{X_i^2}{\big(\sum_{k=1}^n X_k^2\big)^2}$. On en déduit donc que:

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

En rassemblant les morceaux, on obtient:

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

□