1 Exercices

Ci-après, Leb *désigne la mesure de Lebesgue sur* $\mathbb{R}$ *; et* $\mathrm{Leb}^{\otimes n}$ *celle sur* $\mathbb{R}^n$*.*
On écrit i.i.d. pour "indépendantes et identiquement distribuées".
Par convention, les vecteurs sont des vecteurs colonne. Pour $a \in \mathbb{R}^n$*,* a' *désigne la transposée de* a*.*

Exercice 1 (Rappel : transformation de v.a.). Soit un n-échantillon $(X_1, \cdots, X_n)$ du modèle statistique

$$\left(\mathbb{R}^k, \mathcal{B}(\mathbb{R}^k), \left\{p_\theta \cdot \mathrm{Leb}^{\otimes k} \ : \ \theta \in \Theta\right\}\right) .$$

On suppose qu'il existe un ouvert $\mathcal{O}$ de $\mathbb{R}^k$ tel que $\int_{\mathcal{O}} p_\theta \, \mathrm{dLeb}^{\otimes k} = 1$ pour tout $\theta \in \Theta$.

Soit $\phi_\theta : \mathcal{O} \to \mathbb{R}^k$ une application continûment différentiable, injective sur $\mathcal{O}$ et dont le jacobien ne s'annule pas sur $\mathcal{O}$. [1]

1. Sous $p_\theta \cdot \mathrm{Leb}^{\otimes k}$, quelle est la loi de $\phi_\theta(X_i)$?
2. On se place dans le cas $k = 1$, $\theta = (a_1, b_1, \cdots, a_n, b_n)$ et $\Theta = (\mathbb{R} \times \mathbb{R}_\star)^n$. Quel est le modèle statistique induit par $(a_1 + b_1 X_1, \cdots, a_n + b_n X_n)$? Il est d'usage d'appeler a_i le paramètre de *translation* et b_i le paramètre d'*échelle*.

Exercice 2 (Modèle de translation et d'échelle). Soit g une densité par rapport à Leb. On considère le modèle statistique

$$\left(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), \left\{p_{n,\theta} \cdot \mathrm{Leb}^{\otimes n} \ : \ \theta \in \Theta := \mathbb{R} \times \mathbb{R}_+^*\right\}\right)$$

où

$$p_{n,\theta}(x_1, \cdots, x_n) := \sigma^{-n} \prod_{k=1}^{n} g\left(\frac{x_k - \mu}{\sigma}\right), \qquad \theta := (\mu, \sigma).$$

On note $(X_1, \ldots, X_n)$ les variables canoniques : pour tout $i \in \{1, \ldots, n\}$ et $(x_1, \cdots, x_n) \in \mathbb{R}^n$, on a $X_i(x_1, \ldots, x_n) = x_i$.

1. Montrer que sous $p_{n,\theta} \cdot \mathrm{Leb}^{\otimes n}$, les statistiques $(X_1, \ldots, X_n)$ sont i.i.d. et identifier leur loi.
2. Soit $\theta = (\mu, \sigma) \in \Theta$. Montrer que sous $p_{n,\theta} \cdot \mathrm{Leb}^{\otimes n}$, les variables aléatoires réelles

$$\frac{X_i - \mu}{\sigma} , \quad i \in \{1, \ldots, n\}$$

sont i.i.d. de loi de densité g par rapport à Leb.

Supposons que g est une densité gaussienne centrée réduite. On définit les statistiques

$$S_n := \sum_{i=1}^{n} X_i, \qquad K_n := \sum_{k=1}^{n} (X_k - n^{-1} S_n)^2 .$$

1. de façon équivalente, ϕ_θ est un C^1-difféomorphisme de $\mathcal{O}$ sur $\phi_\theta(\mathcal{O})$ i.e. c'est une bijection de $\mathcal{O}$ sur $\phi_\theta(\mathcal{O})$ telle que ϕ_θ et ϕ_θ^{-1} sont de classe C^1.

3. Proposer un estimateur de μ puis de σ^2, utilisant ces deux statistiques. *Commentaires : il n'y a pas unicité d'un estimateur de $T(\theta)$. Dans le cas de l'estimateur de la moyenne, la figure 1 permet de comparer deux stratégies d'estimation : l'une repose sur l'estimateur de la moyenne empirique, et l'autre sur l'estimateur de la médiane. Cette figure affiche le boxplot de* 5000 *réalisations indépendantes de chacun des estimateurs dans le cas $n = 50$ (gauche) et le cas $n = 500$ (droite). Qu'en concluez-vous ? dans le corrigé rédigé, on trouvera les mêmes analyses graphiques pour le cas où g est une densité de Laplace ; pensez-vous que la conclusion sera la-même ?*

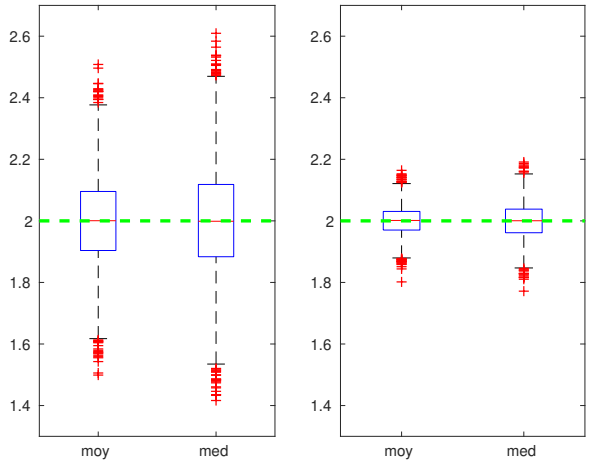


FIGURE 1 – Comparaison de l'estimateur de la moyenne empirique et de celui de la médiane dans le cas où $\mu = 2$ et $\sigma^2 = 1$. Boxplot de 5000 réalisations indépendantes de l'estimateur, dans le cas où $n = 50$ (groupe de gauche) et $n = 500$ (groupe de droite).

4. Déterminer le modèle statistique induit par les statistiques (S_n, K_n). *La démonstration du théorème de Gosset est donnée dans le polycopié de cours (Théorème IV-5-24).*

Exercice 3. Nous cherchons à modéliser la dépendance d'une réponse par rapport à k variables explicatives, sur une population de n individus. On collecte donc

$$(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n) \ ,$$

où $y_i \in \mathbb{R}$ et $\mathbf{x}_i \in \mathbb{R}^k$ sont respectivement la *réponse* et les *variables explicatives* pour le i-ème individu. Considérons le modèle statistique suivant

$$\left(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), \left\{ p_{n,\theta} \cdot \mathrm{Leb}^{\otimes n} \ : \ \theta = (\beta, \sigma) \in \Theta := \mathbb{R}^k \times \mathbb{R}_+^* \right\}\right) ,$$

où

$$p_{n,\theta}(y_1, \cdots, y_n) := \sigma^{-n} \prod_{i=1}^{n} g\left(\frac{y_i - f(\mathbf{x}_i'\beta)}{\sigma}\right) ;$$

g est une densité par rapport Leb, $f : \mathbb{R} \to \mathbb{R}$ une fonction. Pour $i \in \{1, \dots, n\}$ nous notons Y_i la i-ème variable canonique, $Y_i(y_1, \dots, y_n) = y_i$.

1. Montrer que sous $p_{n,\theta}\cdot \mathrm{Leb}^{\otimes n}$, les statistiques $(Y_1,\dots,Y_n)$ sont indépendantes et préciser leurs lois.
2. Montrer que sous $p_{n,\theta}\cdot \mathrm{Leb}^{\otimes n}$ les variables
$$\sigma^{-1}\{Y_i - f(\beta'\mathbf{x}_i)\}, \qquad i\in\{1,\cdots,n\}$$
sont i.i.d. de loi $g\cdot\mathrm{Leb}$.
3. Dans cette question, $k=2$, g est la densité d'une loi gaussienne centrée réduite, $f(\eta)=\eta$ et pour tout $i\in\{1,\cdots,n\}$
$$\mathbf{x}_i = \begin{bmatrix}1\\ x_i\end{bmatrix} \qquad f(\beta'\mathbf{x}_i)=\beta_0+\beta_1 x_i.$$
Proposer un estimateur de β_0 et β_1.

Nous disposons de 150 mesures de concentration d'Ozone (moyennes observées entre 13 :00 et 15 :00 à Roosevelt Island en p.p.m) en fonction de différents facteurs : radiation solaire, vitesse du vent, température. On décide d'explorer la relation entre la concentration d'ozone (considérée comme la variable de réponse) et la radiation solaire (considérée comme la variable explicative).

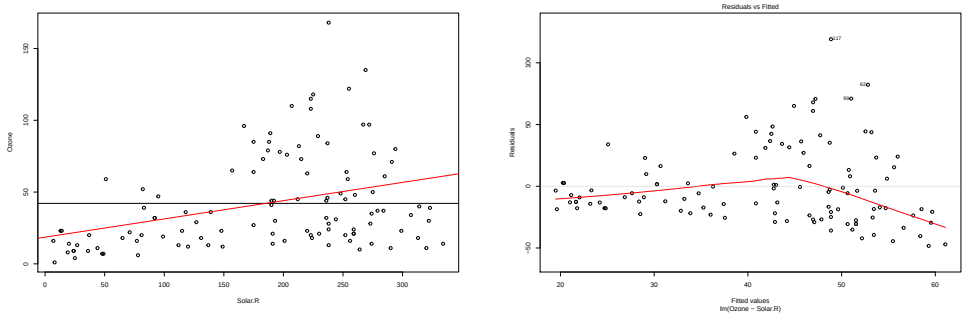


FIGURE 2 – (gauche) Le nuage de 150 points (x_i,y_i) où, pour la mesure i, x_i est la radiation solaire et y_i la concentration d'ozone. On trace aussi, en rouge, la droite de régression $x\mapsto b_0+b_1x$ où b_0,b_1 ont été obtenus comme minimisant la quantité $(\beta_0,\beta_1)\mapsto\sum_{i=1}^{150}(y_i-\beta_0-\beta_1x_i)^2$. La droite horizontale est la droite d'équation $x\mapsto n^{-1}\sum_{k=1}^n y_k$. (droite) Nuage de 150 points $(\widehat{y}_i,e_i)$ où $\widehat{y}_i=b_0+b_1x_i$ et $e_i=y_i-\widehat{y}_i$.

4. Le modèle de régression linéaire vous semble-t-il approprié ? Quelles améliorations du modèle sembleraient souhaitables ?

Exercice 4 (Score de football). Soit $\lambda>0$. On appelle *loi de Poisson* de paramètre λ la loi de densité $\{p_\lambda(k), k\in\mathbb{N}\}$ par rapport à la mesure de comptage μ sur $\mathbb{N}$:
$$p_\lambda(k)=\mathrm{e}^{-\lambda}\frac{\lambda^k}{k!}\,, \quad k\in\mathbb{N}$$

1. Calculer la fonction génératrice des moments de la loi de Poisson de paramètre λ.
2. En déduire la moyenne et la variance d'une loi de Poisson de paramètre λ.

3. Montrer que si X_1 et X_2 sont deux variables indépendantes de loi de Poisson de paramètres $\lambda_1 > 0$ et $\lambda_2 > 0$, alors la variable aléatoire $X_1 + X_2$ suit une loi de Poisson de paramètre $\lambda_1 + \lambda_2$.

Soit $(X_1, \dots, X_n)$ un n-échantillon du modèle

$$(\mathbb{N}, \mathcal{P}(\mathbb{N}), \{p_\lambda \cdot \mu \,:\, \lambda \in \mathbb{R}_+^*\})$$

4. Définir le modèle statistique induit par la statistique $\sum_{i=1}^n X_i$.
5. Proposer une méthode d'estimation du paramètre λ.

On se propose de modéliser le nombre de buts inscrits dans un match de football par une loi de Poisson. On considère tout d'abord que le nombre de buts inscrits par l'équipe locale et l'équipe visiteuse sont deux variables de Poisson indépendantes de loi de Poisson de paramètres différents. On suppose aussi que les résultats des matchs sont indépendants.

6. Construire le modèle statistique associé à l'observation des résultats de n matchs.
7. On a collecté les buts marqués en *premier league* de la saison 2004-2005 à la saison 2008-2009. Le nombre de matchs est de $n = 1900$:

   le nombre de buts marqués est en moyenne 2.523 avec une variance de 2.640,

   le nombre de buts marqués par l'équipe locale est en moyenne 1.468 buts avec une variance de 1.617,

   le nombre de buts marqués par l'équipe visiteuse est en moyenne 1.055 avec une variance de 1.158.

   Proposer un estimateur des intensités λ_1 et λ_2 des deux processus de Poisson introduits pour modéliser le nombre de buts marqués par chacune des deux équipes. Pourquoi l'hypothèse poissonienne est-elle discutable ?

Au lieu d'ajuster une loi de Poisson, il semble plus judicieux dans ce cas de considérer une famille de loi présentant une "sur-dispersion" par rapport à la loi de Poisson (i.e. pour laquelle la variance puisse être plus grande que la moyenne). Etudions plus en détail le nombre de buts marqués par l'équipe locale et l'équipe visiteuse (voir figure 3).

	Observés	Poisson ($\lambda = 1.468$)
0	469	437.7
1	621	642.6
2	456	471.7
3	217	230.8
4	100	84.7
≥ 5	37	32.5
Total	1900	

	Observés	Poisson ($\lambda = 1.055$)
0	692	661.6
1	680	697.9
2	335	368.2
3	131	129.5
4	51	34.1
≥ 5	11	8.7
Total	1900	

FIGURE 3 – Résultats de l'équipe locale (gauche) et de l'équipe visiteuse (droite)

L'analyse de ces résultats suggère que la modélisation poissonnienne sous-estime le nombre de scores nuls et sur-estime en contre-partie les cas où 1 ou 2 buts sont marqués. On considère comme modèle, un mélange de la distribution de Poisson et d'un atome en 0,

$$p_{\pi,\lambda}(k) = (1-\pi)\mathbb{1}_{\{0\}}(k) + \pi \mathrm{e}^{-\lambda}\frac{\lambda^k}{k!}\,,$$

où $\pi \in]0,1[$ est la proportion du mélange.

8. Comment obtenir des réalisations d'une telle loi à partir d'un générateur de v.a. uniforme sur $[0,1]$ et d'un générateur de loi de Poisson de paramètre λ ?
9. Calculer la moyenne et le moment d'ordre 2 de cette distribution.
10. Proposer une méthode d'estimation de π et de λ.

Exercice 5 (Prix d'un actif financier). Soit X une variable aléatoire réelle d'espérance μ et d'écart type σ. Si $\mathbb{E}[|X|^3] < \infty$, on définit le coefficient d'asymétrie comme le moment d'ordre trois de la variable centrée réduite :

$$\gamma_1 = \mathbb{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = \frac{\mu_3}{\mu_2^{3/2}} ,$$

avec μ_i les moments centrés d'ordre i. Si $\mathbb{E}[X^4] < \infty$, on définit son kurtosis non normalisé (coefficient d'aplatissement) comme le moment d'ordre quatre de la variable centrée réduite :

$$\beta_2 = \mathbb{E}\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] = \frac{\mu_4}{{\mu_2}^2} .$$

On définit l'excès de kurtosis comme $\gamma_2 = \beta_2 - 3$.

On observe la suite $p_1, \cdots, p_n$ du prix d'un actif financier à la clôture d'un marché (prix journalier). On modélise ces données comme une réalisation du vecteur aléatoire $(P_1, \cdots, P_n)$. On appelle log-rendement de cet actif la quantité

$$X_k := \log(P_k/P_{k-1}) .$$

Un modèle couramment utilisé (associé à la théorie proposée par F. Black, M. Scholes et R. Merton, prix Nobel 1997) consiste à supposer que (i) les log-rendements $\{X_i, i \geq 1\}$ sont i.i.d. et (ii) ils suivent une distribution gaussienne (dont la moyenne μ et la variance σ^2 sont inconnues ; l'analyse statistique peut consister à estimer ces quantités).

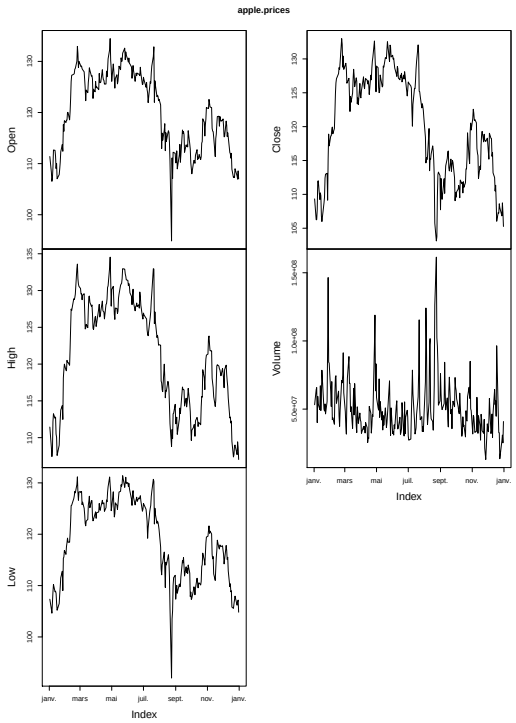
On souhaite estimer μ, σ^2 à partir de mesures des log-rendements.

1. Proposer un modèle statistique des log-rendements.
2. Proposer un estimateur de la moyenne μ et de la variance σ^2.
3. On superpose l'histogramme des observations et la densité d'une loi gaussienne dont les paramètres sont égaux à ceux estimés (voir figure). Peut-on être satisfait de ce modèle ? Qu'observe-t-on ?
4. Lorsque X est une variable aléatoire gaussienne de moyenne μ et de variance σ^2,
   a) montrer que le coefficient d'asymétrie est nul.
   b) calculer $\mathbb{E}[\mathrm{e}^{tX}]$ pour tout $t \in \mathbb{R}$. En identifiant les premiers termes du développement de $\log \mathbb{E}[\mathrm{e}^{tX}]$, montrer que l'excès de kurtosis est nul.
5. Proposer un estimateur empirique du coefficient d'asymétrie et du coefficient d'excès de kurtosis.
6. En évaluant ces estimateurs sur la série des log-rendements, nous obtenons -0.0707 pour l'asymétrie et 1.274035 pour l'excès de kurtosis. Le modèle retenu vous semble-t-il acceptable ?

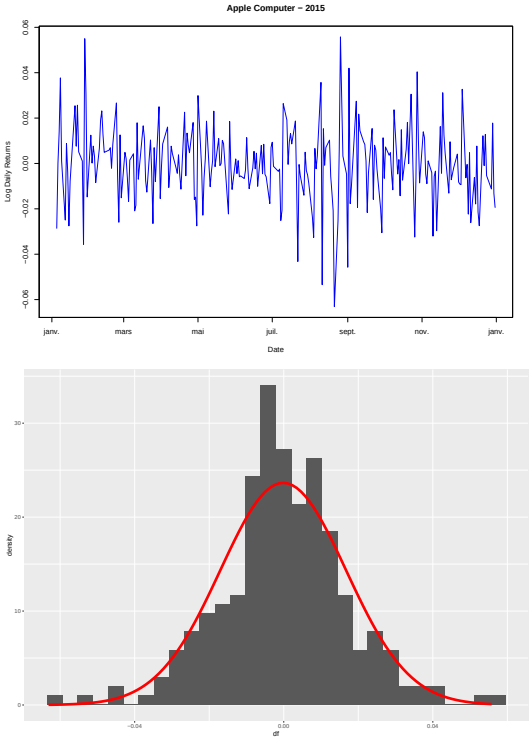
7. Pour modéliser les log-rendements, R. Engle (Prix Nobel d'Economie 2003) a proposé le modèle ARCH(1) :

$$X_k = \sqrt{\alpha_0 + \alpha_1 X_{k-1}^2} Z_k \,, \quad X_0 = 0 \,,$$

où $\alpha_0 > 0$ et $\alpha_1 \geq 0$ et $\{Z_k\}_{k=1}^n$ est une suite i.i.d. de variables aléatoires gaussiennes centrées réduites. Définir le modèle statistique associé.



Mesures, dont le prix à la clôture



Log-rendements (haut) et ajustement d'une gaussienne (bas)

Exercice 6 (Sur l'identifiabilité des lois de mélange Gaussien et Poisson). On appelle simplexe des probabilités l'ensemble

$$\mathsf{S} := \left\{ (\alpha_1, \ldots, \alpha_K) \in (\mathbb{R}_+)^K \,:\, \sum_{i=1}^K \alpha_i = 1 \right\}.$$

On pose d'autre part

$$\mathsf{M} := \left\{ \mu = (\mu_1, \ldots, \mu_K) \in \mathbb{R}^K \,:\, \mu_1 < \mu_2 < \ldots < \mu_K \right\}.$$

1. Démontrer que pour tout $(\mu_1, \ldots, \mu_K) \in \mathsf{M}$, la famille de fonctions $\left(t \longmapsto \mathrm{e}^{\mathrm{i}t\mu_j}\right)_{1 \leq j \leq K}$ est une famille libre.

Pour tout $\alpha = (\alpha_1, \ldots, \alpha_K) \in \mathsf{S}$ et $\mu = (\mu_1, \ldots, \mu_K) \in \mathsf{M}$, on considère la loi de densité par rapport à la mesure de Lebesgue

$$f_{\alpha,\mu}(x) = \frac{1}{\sqrt{2\pi}} \sum_{k=1}^K \alpha_k \, \mathrm{e}^{-(x-\mu_k)^2/2}.$$

2. Démontrer que le modèle statistique de mélange gaussien $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \{f_{\alpha,\mu} \cdot \text{Leb} : (\alpha, \mu) \in \mathsf{S} \times \mathsf{M}\})$ est identifiable. On rappelle pour cela la fonction caractéristique d'une variable aléatoire X gaussienne centrée réduite :
$$\mathbb{E}[e^{itX}] = e^{-t^2/2}.$$

La loi de Poisson de paramètre μ est la loi de densité $\{p_\mu(n) := \mathrm{e}^{-\mu}\frac{\mu^n}{n!}, n \in \mathbb{N}\}$ par rapport à la mesure de comptage ν sur $\mathbb{N}$. Sa fonction caractéristique est donnée par

$$\Phi_{p_\mu}(t) = \exp[\mu(\mathrm{e}^{it} - 1)].$$

Pour tout $\alpha \in \mathsf{S}$ et $\mu = (\mu_1, \ldots, \mu_K) \in \mathsf{M}$, on définit la densité de mélange Poissonien

$$q_{\alpha,\mu}(n) = \sum_{k=1}^{K} \alpha_k \; p_{\mu_k}(n).$$

3. Démontrer que le modèle statistique *de mélange poissonnien* : $(\mathbb{N}, \mathcal{P}(\mathbb{N}), \{q_{\alpha,\mu} \cdot \nu, (\alpha, \mu) \in \mathsf{S} \times \mathsf{M}\})$ est identifiable.