Dans tout l'énoncé,

— Les vecteurs sont des vecteurs colonnes.

— Leb est la mesure de Lebesque sur $\mathbb{R}$ et $\text{Leb}^{\otimes n}$ est la mesure de Lebesgue sur $\mathbb{R}^n$.

— Pour $p > 0$ et $\lambda > 0$, on appelle loi Gamma(p, λ), la loi de densité donnée par [**attention : convention différente du polycopié**] :

$$f_{p,\lambda}(x) := \frac{\lambda^{-p}}{\Gamma(p)} \exp(-x/\lambda)\ x^{p-1}\ , \quad x > 0\ .$$

On note Gcdf_p et Gqf_p la fonction de répartition et la fonction quantile de la Gamma$(p, 1)$.

Nous rappelons

— Si la variable aléatoire Y est distribuée suivant une loi Gamma(p, λ) avec $p > 0$, $\lambda > 0$, alors Y/λ est distribuée suivant une loi Gamma$(p, 1)$.

— Si la variable aléatoire Y est distribuée suivant une loi Gamma(p, λ) avec $p > 0$, $\lambda > 0$ alors $\mathbb{E}[Y] = p\lambda$ et $\text{Var}(Y) = p\lambda^2$.

— La loi du χ^2 centré à n degrés de liberté est une loi Gamma$(n/2, 2)$.

— Soit $(X_1, \ldots, X_n)$ n variables aléatoires indépendantes, et pour tout $i \in \{1, \ldots, n\}$, la loi de X_i est la loi Gamma(p_i, λ) pour $p_i > 0$, $\lambda > 0$. Alors, $\sum_{i=1}^n X_i$ est distribuée suivant une loi Gamma$(\sum_{i=1}^n p_i, \lambda)$.

Exercice 1. Soit $Z = (X_1, \ldots, X_n)$ un n-échantillon du modèle $(\mathbb{R}_+, \mathcal{B}(\mathbb{R}_+), \{p_\theta.\text{dLeb}, \theta \in \Theta := \mathbb{R}_+^*\})$ où p_θ est la densité de Weibull définie par

$$x \mapsto p_\theta(x) = \frac{c}{\theta}\, x^{c-1} \mathrm{e}^{-x^c/\theta} \mathbb{1}_{\mathbb{R}_+}(x)\ , \quad c \text{ est une constante positive connue}$$

Soit $\theta_0 > 0$. On considère le test

$$H_0 :\ \theta \le \theta_0, \quad \text{contre} \quad H_1 :\ \theta > \theta_0\ .$$

1. Montrer que, pour tout $\theta \in \Theta$, $(X_i)^c$ est, sous $\mathbb{P}_\theta$, distribuée suivant une loi exponentielle de paramètre θ, de densité $q_\theta(y) := \theta^{-1}\mathrm{e}^{-y/\theta}\mathbb{1}_{\mathbb{R}_+}(y)$ [qui est aussi Gamma$(1, \theta)$].
2. Montrer que la famille $\{p_\theta(\cdot), \theta \in \Theta\}$ est à rapport de vraisemblance monotone.
3. Soit $\alpha \in]0, 1[$. Déterminer un test U. P. P.(α) pour ces hypothèses. On explicitera le calcul des constantes dans la définition de la fonction critique du test en fonction de la fonction quantile $\text{Gqf}_n(\cdot)$.
4. Montrer que la fonction puissance du test est donnée par

$$\theta \mapsto 1 - \text{Gcdf}_n\left(\theta_0\, \text{Gqf}_n(1-\alpha)/\theta\right)$$

Exercice 2. Soit $(X_1, \ldots, X_n)$ un n-échantillon du modèle $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \{p_\theta.\text{dLeb}, \theta \in \Theta\})$, où $\theta := (\mu, \sigma^2) \in \Theta := \mathbb{R} \times \mathbb{R}_+^*$, et p_θ est la densité de probabilité d'une loi Gaussienne de moyenne μ et variance σ^2. Nous considérons les estimateurs de μ^2 donnés par

$$T_{1,n} := \left(\bar{X}_n\right)^2 \quad \text{où} \quad \bar{X}_n := \frac{1}{n}\sum_{i=1}^n X_i$$

$$T_{2,n} := \left(\bar{X}_n\right)^2 - \frac{1}{n}S_n^2 \quad \text{où} \quad S_n^2 := \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2\ .$$

On rappelle que Y est une variable aléatoire gaussienne centrée réduite, alors $\mathbb{E}[Y^4] = 3$.

1. Montrer que, pour tout $\theta \in \Theta$, sous $\mathbb{P}_{n,\theta}$
   (a) la variable aléatoire $\bar{X}_n$ suit une loi gaussienne dont on déterminera la moyenne et la variance.
   (b) la variable S_n^2 est indépendante de $\bar{X}_n$ et $(n-1)S_n^2/\sigma^2$ est distribuée suivant une loi Gamma dont on précisera les paramètres.
   (c) Montrer que $\mathbb{E}_\theta[S_n^2] = \sigma^2$ and $\mathbb{E}_\theta[S_n^4] = \sigma^4(1 + 2/(n-1))$.
2. Calculer le biais de l'estimateur et le risque quadratique de l'estimateur $T_{1,n}$ de μ^2.
3. Calculer le biais et le risque quadratique de l'estimateur $T_{2,n}$ de μ^2.
4. L'estimateur $T_{1,n}$ de μ^2 est-il admissible pour le risque quadratique ?
5. Montrer que les suites d'estimateurs $\{T_{1,n}, n \geq 2\}$ et $\{T_{2,n}, n \geq 2\}$ sont consistantes.
6. Montrer que pour tout $\theta = (\mu, \sigma^2) \in \mathbb{R}^* \times \mathbb{R}_+^*$ (i.e. $\mu \neq 0$), $\sqrt{n}(T_{1,n} - \mu^2)$ converge sous $\mathbb{P}_{n,\theta}$ vers une loi gaussienne dont on déterminera la variance. Même question pour $\sqrt{n}(T_{2,n} - \mu^2)$.
7. Montrer que pour tout θ tel que $\mu = 0$, $nT_{1,n}$ converge en loi sous $\mathbb{P}_{n,\theta}$ vers une loi que l'on caractérisera. Même question pour $nT_{2,n}$. Commenter.

Exercice 3. Soit $Z := (X_1, \ldots, X_n)$ un n-échantillon du modèle statistique du modèle exponentiel translaté $\mathcal{E}(a, \lambda)$ de densité par rapport à la mesure de Lebesgue

$$\left(\mathbb{R}; \mathcal{B}(\mathbb{R}); \{f_\theta \cdot \mathrm{Leb}, \theta := (a, \lambda) \in \Theta := \mathbb{R} \times \mathbb{R}_+^*\}\right), \quad \text{où} \quad f_\theta(x) := \lambda^{-1} \mathrm{e}^{-(x-a)/\lambda} \mathbb{1}_{[a,\infty[}(x)$$

Nous notons, pour $\theta \in \Theta$,

$$p_\theta(x_1, \cdots, x_n) := \prod_{k=1}^n f_\theta(x_k),$$

et $\mathbb{P}_\theta := p_\theta \cdot \mathrm{Leb}^{\otimes n}$. Soit $X_{1:n} \leq X_{2:n} \leq \cdots \leq X_{n:n}$ les statistiques d'ordre de l'échantillon ; puisque les variables aléatoires $(X_k)_{k=1}^n$ ont des lois à densité par rapport à la mesure de Puisque les variables aléatoires $(X_k)_{k=1}^n$ ont des lois à densité par rapport à la mesure de Lebesgue, on a $\mathbb{P}_\theta(X_i = X_j) = 0$ pour tout $i \neq j$ et $\theta \in \Theta$. Par suite, il existe, $\mathbb{P}_\theta$-p.s. une unique permutation, que nous notons $\sigma(Z) = [\sigma_1(Z), \ldots, \sigma_n(Z)]$ de $\{1, \cdots, n\}$ telle que

$$X_{k:n} = X_{\sigma_k(Z)}, \qquad X_{\sigma_1(Z)} < X_{\sigma_2(Z)} < \cdots < X_{\sigma_n}(Z), \quad \mathbb{P}_\theta - \text{p.s.}$$

Nous notons $\mathcal{S}_n$ l'ensemble des permutations de l'ensemble $\{1, \ldots, n\}$.

1. Montrer que pour toute fonction h mesurable bornée

$$\mathbb{E}_\theta \left[h\left(X_{1:n}, \cdots, X_{n:n}\right)\right] = \sum_{\tau \in \mathcal{S}_n} \mathbb{E}_\theta \left[h\left(X_{\tau_1}, \cdots, X_{\tau_n}\right) \mathbb{1}_{\{X_{\tau_1} < X_{\tau_2} < \cdots < X_{\tau_n}\}}\right] \tag{1}$$

2. En déduire que sous $\mathbb{P}_\theta$, la loi jointe des statistiques d'ordre $X_{1:n}, X_{2:n}, \ldots, X_{n:n}$ admet une densité par rapport à $\mathrm{Leb}^{\otimes n}$, densité donnée par

$$(y_1, \cdots, y_n) \mapsto \varphi(\theta; y_1,\ y_2,\ \ldots,\ y_n) := n!\, p_\theta(y_1, \cdots, y_n)\ \mathbb{1}_{\{a \leq y_1 < y_2 < \cdots < y_n\}}\ .$$

On considère la transformation

$$U_1 := nX_{1:n}, \qquad U_k := (n-k+1)(X_{k:n} - X_{(k-1):n}) \;, \quad \text{pour } k \in \{2, \dots, n\} \;; \tag{2}$$

Remarquons que pour tout $(u_1, \dots, u_n) \in \mathbb{R}^n$,

$$\sum_{i=1}^{n} u_i = nu_1 + \sum_{i=2}^{n} (n-i+1)(u_i - u_{i-1}) \;. \tag{3}$$

3. Montrer que, sous $\mathbb{P}_\theta$, les variables aléatoires $(U_1, \dots, U_n)$ sont indépendantes. Déterminer également la loi de chacune de ces variables.
4. Démontrer que sous $\mathbb{P}_\theta$, la statistique $T_1 := \sum_{i=1}^{n} (X_{i:n} - X_{1:n})$ est indépendante de $X_{1:n}$ et suit une loi Gamma$(n-1, \lambda)$.
5. Construire, en utilisant T_1, un intervalle de confiance pour λ avec une probabilité de couverture de $1-\alpha$, où $\alpha \in]0, 1[$.
6. Établir que sous $\mathbb{P}_\theta$, la statistique $T_2 := X_{1:n}$ suit une loi $\mathcal{E}(a, \lambda/n)$.
7. Justifier *sans calcul* que la fonction $a \mapsto F_n(a) := (T_2 - a)/\{(n-1)^{-1} T_1\}$ est pivotale pour le paramètre θ.
8. **Question bonus !** Prouver que, pour tout $\theta \in \Theta$, la loi de $F_n(a)$ par rapport à la mesure de Lebesgue est donnée par l'équation suivante :

$$t \mapsto n \left(1 + \frac{nt}{n-1}\right)^{-n} \mathbb{1}_{[0,\infty[}(t) \;. \tag{4}$$

9. En se servant de la fonction pivotale $F_n(a)$, construire un intervalle de confiance pour a avec une probabilité de couverture de $1-\alpha$.

Exercice 4. On cherche à expliquer une variable quantitative Y_i (réponse) par une régression linéaire à un facteur $[1, x_i]^\top$, $i \in \{1, \dots, n\}$, $x_i \in \mathbb{R}$. On pose $\theta := (\boldsymbol{\beta}, \sigma^2) \in \Theta := \mathbb{R} \times \mathbb{R}^* \times \mathbb{R}_+^*$, avec $\boldsymbol{\beta} := [\beta_0, \beta_1]^\top$. Nous posons $\mathbf{Y} := [Y_1, \dots, Y_n]^\top$, $\mathbf{y} := [y_1, \dots, y_n]^\top$. Nous supposons que, pour tout $\theta \in \Theta$, $\mathbf{Y}$ est un vecteur Gaussien, de moyenne $[\beta_0 + \beta_1 x_1, \dots, \beta_0 + \beta_1 x_n]^\top$ et de covariance $\sigma^2 \mathrm{I}_n$, i.e. la densité de $\mathbf{Y}$ par rapport à la mesure de Lebesgue sur $\mathbb{R}^n$ est

$$p_\theta(\mathbf{y}) := \frac{1}{\sqrt{2\pi\sigma^2}^n} \exp\left(-\frac{1}{2} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2\right)$$

Posons $\mathbf{x} := [x_1, \dots, x_n]^\top$, $\mathbf{1} := [1, \dots, 1]^\top \in \mathbb{R}^n$, et $\mathbf{X} := [\mathbf{1}, \mathbf{x}]$. Avec ces notations, nous pouvons écrire de façon équivalente

$$p_\theta(\mathbf{y}) = \frac{1}{\sqrt{2\pi\sigma^2}^n} \exp\left(-\frac{1}{2} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2\right)$$

Nous supposons que la matrice $\mathbf{X}$, de taille $n \times 2$, est de rang 2 et nous posons $H := \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$ le projecteur orthogonal sur Vect$(\mathbf{X})$.

Nous notons

$$\bar{Y} := \frac{1}{n} \sum_{i=1}^{n} Y_i = \frac{\mathbf{1}^\top \mathbf{Y}}{\|\mathbf{1}\|^2} \quad \text{et} \quad \bar{x} := \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{\mathbf{1}^\top \mathbf{x}}{\|\mathbf{1}\|^2} \;.$$

1. Montrer que les estimateurs du maximum de vraisemblance $\hat{\boldsymbol{\beta}} := [\hat{\beta}_0, \hat{\beta}_1]^\top$ et $\hat{\sigma}^2$ de (β_0, β_1) et σ^2 sont donnés par

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} := (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y} = \begin{bmatrix} \bar{Y} - \hat{\beta}_1 \bar{x} \\ \frac{\sum_{i=1}^n Y_i (x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} \end{bmatrix} \tag{5}$$

$$\hat{\sigma}^2 := \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = \frac{1}{n} \|(\mathrm{I}_n - H)\mathbf{Y}\|^2 \tag{6}$$

Nous remarquons que $\mathbb{P}_\theta$-p.s., $\hat{\beta}_1 \neq 0$.

2. Démontrer que, sous $\mathbb{P}_{\boldsymbol{\beta},\sigma^2}$, $\hat{\boldsymbol{\beta}}$ est un vecteur gaussien dont on précisera la moyenne et la covariance.
3. Montrer que les statistiques $\hat{\sigma}^2$ et $\hat{\boldsymbol{\beta}}$ sont indépendantes. Déterminer la distribution de la statistique $\hat{\sigma}^2$.

Nous posons, $S^2 := (n/(n-2))\hat{\sigma}^2$ et pour $\boldsymbol{\beta} := [\beta_0, \beta_1]^\top \in \mathbb{R} \times \mathbb{R}^*$: $\phi := -\beta_0/\beta_1$, et $\delta := -\phi + \bar{x}$.

4. Montrer que, pour tout $\theta \in \Theta$,

$$\mathbb{E}_\theta \left[\bar{Y} - \delta \hat{\beta}_1 \right] = 0 \,.$$

5. Montrer que, pour tout $\theta \in \Theta$,

$$\mathrm{Cov}_\theta(\bar{Y}, \hat{\beta}_1) = 0 \,.$$

6. Montrer que, pour tout $\theta \in \Theta$,

$$\mathrm{Var}_\theta \left(\bar{Y} - \delta \hat{\beta}_1 \right) = \sigma^2 w(\delta) \quad \text{où} \quad w(\delta) := \frac{1}{n} + \frac{\delta^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \,.$$

7. Montrer que pour tout $\theta \in \Theta$, sous $\mathbb{P}_\theta$, $\bar{Y} - \delta \hat{\beta}_1$ est distribué suivant une loi normale centrée de variance $\sigma^2 w(\delta)$ et est indépendante de S^2.

Nous posons

$$T(\delta) := \frac{\bar{Y} - \delta \hat{\beta}_1}{S \sqrt{w(\delta)}}$$

8. Montrer que pour tout $\theta \in \Theta$, sous $\mathbb{P}_\theta$, $T^2(\delta)$ est distribuée suivant une loi de Fisher à $(1, n-2)$ degrés de liberté.
9. En déduire une région de confiance pour δ de probabilité de couverture $1-\alpha$, $\alpha \in]0,1[$, de la forme $A(\alpha)\delta^2 + B(\alpha)\delta + C(\alpha) \leq 0$ [on déterminera $A(\alpha)$, $B(\alpha)$, et $C(\alpha)$]. Sous quelle condition cette région de confiance est-elle un intervalle ?

Exercice 5. Dans cet exercice, nous considérons des observations i.i.d. $(X_i, Y_i)_{1 \leq i \leq n}$ où Y_i prend des valeurs dans $\{-1, 1\}$ et X_i prend des valeurs dans $\mathbb{R}^d$. Pour un x donné dans $\mathbb{R}^d$, nous considérerons le classificateur suivant : $\hat{h}_n(x) = Y_{\phi_n(x)}$ où

$$\phi_n(x) = \arg \min_{i \in [1:n]} \|x - X_i\|$$

En d'autres termes, $\phi_n(x)$ est l'indice du voisin le plus proche de x parmi l'ensemble de données $\{X_i; i \in [1:n]\}$. Dans tout l'exercice, nous considérons une paire de variables aléatoires (X, Y) ayant la même loi que (X_i, Y_i) pour tout $i \geq 1$. Et pour éviter toute ambiguïté dans la définition de l'indice $\phi_n(X)$, nous supposons qu'avec probabilité 1, tous les $\|X - X_i\|$ pour tout $i \in \mathbb{N}$ sont strictement différents.

Rappelons que le classificateur optimal de Bayes est défini par

$$h^\star(X) = \begin{cases} 1 & \text{si } \mathbb{P}(Y=1|X) > \mathbb{P}(Y=-1|X) \\ -1 & \text{sinon} \end{cases}$$

et rappelons qu'en définissant $\eta(X) = \mathbb{P}(Y=1|X)$ et $r^\star(X) = \min(\eta(X), 1-\eta(X))$, nous avons

$$R^\star = \mathbb{P}(Y \neq h^\star(X)) = \mathbb{E}[r^\star(X)] \leq \mathbb{P}(Y \neq h(X))$$

pour tout autre classificateur $h : \mathbb{R}^d \to \{-1, 1\}$. Le but de cet exercice est de montrer la borne

$$R^\star \leq \lim_{n\to\infty} \mathbb{P}(Y \neq \hat{h}_n(X)) \leq 2R^\star(1-R^\star)$$

Définissons $X_{(n)} = X_{\phi_n(X)}$ le voisin le plus proche de X parmi l'ensemble $\{(X_i); i \in [1:n]\}$. Nous admettons que,

— la suite $\{X_{(n)},\ (n) \in \mathbb{N}\}$ *converge en probabilité* vers X :

$$X_{(n)} \overset{\mathbb{P}-\text{prob}}{\longrightarrow} X\ ,$$

— la fonction η est continue,

1. Montrez que, pour tout $x \in \mathbb{R}^d$, $r^\star(x)(1-r^\star(x)) = \eta(x)(1-\eta(x))$.
2. Montrez que pour tout $i \in [1:n]$, nous avons presque sûrement,

$$\mathbb{P}(Y=-1, Y_i=1, \phi_n(X)=i|X, X_{1:n}) = (1-\eta(X))\eta(X_i)\mathbb{1}_{\{i\}}(\phi_n(X))$$

3. En notant que, $\eta(X_{(n)}) = \sum_{i=1}^n \eta(X_i)\mathbb{1}_{\{i\}}(\phi_n(X))$, déduisez que, presque sûrement,

$$\mathbb{P}(Y=-1, \hat{h}_n(X)=1|X, X_{1:n}) = (1-\eta(X))\eta(X_{(n)}),$$

4. De la même manière, montrez que

$$\mathbb{P}(Y=1, \hat{h}_n(X)=-1|X, X_{1:n}) = \eta(X)(1-\eta(X_{(n)})),$$

5. Montrez que $\lim_{n\to\infty} \mathbb{E}[\eta(X_{(n)})] = \mathbb{E}[\eta(X)]$.
6. Montrez que :

$$\lim_{n\to\infty} \mathbb{P}(Y \neq \hat{h}_n(X)) = 2\mathbb{E}[r^\star(X)(1-r^\star(X))]$$

7. Déduisez que

$$\lim_{n\to\infty} \mathbb{P}(Y \neq \hat{h}_n(X)) \leq 2R^\star(1-R^\star)$$

et conclure.