Exercices de statistiques mathématiques

Guillaume Lecué

5 septembre 2016

Table des matières

1 Rappels de probabilités

****

Exercice 1.1 (Théorème de la limite centrale)
Soit $(X_n)_n$ une suite de variables aléatoires i.i.d. centrées de variance $\sigma^2 > 1$. Soit

$$Z_n = \frac{1}{\sigma\sqrt{n}} \sum_{j=1}^{n} X_j \,.$$

Par le théorème de la limite centrale, cette variable converge en loi vers la loi normale centrée réduite, c'est-à-dire, pour tout $t \in \mathbb{R}$, on a $\lim_{n\to+\infty} \mathbb{E}[e^{itZ_n}] = e^{-\frac{t^2}{2}}$. L'objet de cet exercice est de montrer que la suite Z_n ne peut pas converger en probabilité.

1. **Calculer la fonction caractéristique de $Z_{2n} - Z_n$ et montrer que cette différence converge en loi.**
2. **En étudiant $\mathbb{P}(|Z_{2n} - Z_n| \geq \epsilon)$, montrer que Z_n ne converge pas en probabilité.**

****
****

Exercice 1.2 (Lemme de Slutsky)

1. **Donner un exemple de suites (X_n) et (Y_n) telles que $X_n \overset{\text{loi}}{\to} X$ et $Y_n \overset{\text{loi}}{\to} Y$, mais $X_n + Y_n$ ne converge pas en loi vers $X + Y$.**
2. **Soient (X_n), (Y_n) deux suites de variables aléatoires réelles, X et Y des variables aléatoires réelles, telles que**
   (i) $X_n \overset{\text{loi}}{\to} X$ et $Y_n \overset{\mathbb{P}}{\to} Y$,
   (ii) Y est indépendante de (X_n) et X.
   Montrer que le couple (X_n, Y_n) converge en loi vers (X, Y).
3. **En déduire que si (X_n) et (Y_n) sont deux suites de variables aléatoires réelles telles que (X_n) converge en loi vers une limite X et (Y_n) converge en probabilité vers une constante c, alors $(X_n + Y_n)$ converge en loi vers $X + c$ et $(X_n Y_n)$ converge en loi vers cX.**

****
****

Exercice 1.3 (Convergence dans L^p)

Soit (X_n) une suite de variables aléatoires réelles bornées par une même constante. Montrer que si (X_n) converge en probabilité, alors X_n converge dans L^p pour tout $p \geq 1$.

****
****

Exercice 1.4 (Loi conditionnelle)

Soit X une variable aléatoire qui suit une loi Gamma $(2, \lambda)$ de densité

$$f(x) = \lambda^2 x e^{-\lambda x} \mathbb{1}_{[0,+\infty)}(x)$$

et soit Y une variable aléatoire dont la loi conditionnelle à $X = x$ est uniforme sur $[0, x]$.

1. **Donner la loi jointe de (X, Y).**
2. **Donner la loi marginale de Y et montrer que Y est indépendant de $X - Y$.**

****
****

Exercice 1.5 (Estimateur de la variance)

Soient $X_1, \ldots, X_n$ des variables aléatoires i.i.d., $X_i \sim f(\cdot - \theta)$, où f est une densité de probabilité sur $\mathbb{R}$ symétrique dont on note $\mu_k = \int_{\mathbb{R}} x^k f(x)\, dx$ les moments d'ordre $k = 2$ et $k = 4$. On note $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. Montrer que l'estimateur $\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2$ de la variance des X_i vérifie un théorème central limite.

Indication : **on montrera d'abord que l'on peut se ramener au cas où $\theta = 0$, puis on exprimera l'estimateur comme une transformation de $S_n = \frac{1}{n}\sum_{i=1}^n X_i^2$ et de $\bar{X}_n$.**

****
****

Exercice 1.6 (Stabilisation de la variance)

On dispose d'un échantillon $X_1, \ldots, X_n$ i.i.d. de loi de Bernoulli de paramètre $0 < \theta < 1$.

1. **On note $\bar{X}_n$ la moyenne empirique des X_i. Appliquer la loi forte des grands nombres et le TCL dans ce modèle.**
2. **Cherchez une fonction g telle que $\sqrt{n}(g(\bar{X}_n) - g(\theta))$ converge en loi vers Z de loi $\mathcal{N}(0,1)$.**
3. **On note z_α le quantile d'ordre $1 - \alpha/2$ de la loi normale standard. En déduire un intervalle de confiance $\hat{I}_{n,\alpha}$ fonction de $z_\alpha, n, \bar{X}_n$ tel que $\lim_{n\to\infty} \mathbb{P}(\theta \in \hat{I}_{n,\alpha}) = 1 - \alpha$.**

****
****

Exercice 1.7 (Les statistiques d'ordre)

Soient $X_1, \ldots, X_n$ des variables aléatoires i.i.d. de fonction de répartition F. On suppose que F admet une densité f par rapport à la mesure de Lebesgue. On note $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$ les variables aléatoires $X_1, \ldots, X_n$ réordonnées par ordre croissant.

1. **Donner l'expression de la loi de la statistique d'ordre $(X_{(1)}, \ldots, X_{(n)})$ en fonction de f.**
2. **Déterminer la fonction de répartition $F_k(x)$ puis la densité $f_k(x)$ de $X_{(k)}$.**
3. **Sans utiliser les résultats des questions précédentes, calculer les fonctions de répartition de $X_{(1)}$, $X_{(n)}$, du couple $(X_{(1)}, X_{(n)})$ et la loi de la statistique $W = X_{(n)} - X_{(1)}$ (on appelle W *étendue*). Les variables $X_{(1)}$ et $X_{(n)}$ sont–elles indépendantes ?**

****
****

Exercice 1.8 (Durée de vie)

Un système fonctionne en utilisant deux machines de types différents. Les durées de vie X_1 et X_2 des deux machines suivent des lois exponentielles de paramètres λ_1 et λ_2. Les variables aléatoires X_1 et X_2 sont supposées indépendantes.

1. **Montrer que**
$$X \overset{Loi}{=} \mathcal{E}(\lambda) \Leftrightarrow \forall x > 0, \mathbb{P}(X > x) = \mathbf{exp}(-\lambda x).$$
2. **Calculer la probabilité pour que le système ne tombe pas en panne avant la date t. En déduire la loi de la durée de vie Z du système. Calculer la probabilité pour que la panne du système soit due à une défaillance de la machine 1.**

3. **Soit $I = 1$ si la panne du système est due à une défaillance de la machine 1, $I = 0$ sinon. Calculer $\mathbb{P}(Z > t; I = \delta)$, pour tout $t \geq 0$ et $\delta \in \{0, 1\}$. En déduire que Z et I sont indépendantes.**

4. **On dispose de n systèmes identiques et fonctionnant indépendamment les uns des autres dont on observe les durées de vie $Z_1, \ldots, Z_n$.**
   (a) Écrire le modèle statistique correspondant. A-t-on suffisamment d'information pour estimer λ_1 et λ_2 ?
   (b) Si on observe à la fois les durées de vie des systèmes et la cause de la défaillance (machine 1 ou 2), a-t-on alors suffisamment d'information pour estimer λ_1 et λ_2 ?

5. **On considère maintenant un seul système utilisant une machine de type 1 et une machine de type 2, mais on suppose que l'on dispose d'un stock de n_1 machines de type 1, de durées de vie $X_1^1, \ldots, X_1^{n_1}$ et d'un stock de n_2 machines de type 2, de durées de vie $X_2^1, \ldots, X_2^{n_2}$. Quand une machine tombe en panne, on la remplace par une machine du même type, tant que le stock de machines de ce type n'est pas épuisé. Quand cela arrive, on dit que le système lui-même est en panne. On note toujours Z la durée de vie du système. Le cas $n_1 = n_2 = 1$ correspond donc aux trois premières questions.**
   (a) Montrer que la densité de la somme U de k variables indépendantes qui suivent une loi exponentielle de même paramètre λ s'écrit, pour $x \geq 0$:

$$f_U(x) = \frac{\lambda^k}{(k-1)!} x^{k-1} \mathbf{exp}(-\lambda x).$$

   (b) Écrire Z en fonction des X_i^j et en déduire $\mathbb{P}(Z \geq t)$ en fonction n_1, n_2, λ_1, λ_2 et t.

****
****

Exercice 1.9 (Lemme de Fatou)
si (f_n) est une suite de fonctions measurables alors

$$\int \text{liminf}_n f_n \leq \text{liminf}_n \int f_n.$$

En déduire que si (A_n) est une suite d'événements alors

$$\text{limsup}_{\mathbf{n}} \mathbb{P}(\mathbf{A_n}) \leq \mathbb{P}(\text{limsup}_{\mathbf{n}} \mathbf{A_n}),$$

où on rappelle que $\text{limsup}_n A_n = \cap_N \cup_{n \geq N} A_n$.

****
****

Exercice 1.10 (la loi du $0-1$ de Kolmogorov)
Soit (σ_n) une suite de tribus indépendantes. La tribu asymptotique est $\sigma_\infty = \cap_n \sigma\Big(\cup_{p \geq n} \sigma_p\Big)$. La loi du $0-1$ de Kolmogorov dit que pour tout $A \in \sigma_\infty$, $\mathbb{P}[A] \in \{0, 1\}$.

****
****

Exercice 1.11 (convergence en loi vers une constante)
La convergence en loi vers une constante implique la convergence en proba : On suppose $X_n \rightsquigarrow c$ alors (X_n) converge en probabilité vers c.

****
****

Exercice 1.12 (lemmes de Borel-Cantelli)

1. **Le premier lemme de Borel-Cantelli dit que si (A_n) est une suite d'événements telle que $\sum_n \mathbb{P}[A_n] < \infty$ alors $\mathbb{P}[\limsup_n A_n] = 0$.**
2. **Le deuxième lemme de Borel-Cantelli dit que si (A_n) est une suite d'événements indépendants tels que $\sum_n \mathbb{P}[A_n] = \infty$ alors $\mathbb{P}[\limsup_n A_n] = 1$.**

****
****

Exercice 1.13 (L'asymptotique normalité implique la converge en probabilité)
Soit (r_n) une suite de réels positifs tendant vers $+\infty$. Soit (ζ_n) une suite de v.a.r. telle que $r_n(\zeta_n - \mu) \rightsquigarrow \zeta$. Alors (ζ_n) converge en probabilité vers μ.

****
****

Exercice 1.14 (quartile)
Soit la loi de probabilité de densité $f(x) = 2xI\{0 \leq x \leq 1\}$.

1. **Trouver les quartiles (y compris la médiane) de cette loi.**
2. **Considérons un échantillon i.i.d. $(X_1, \ldots, X_n)$ de cette loi. Soit $\widehat{F}_n$ la fonction de répartition empirique associée. Donner la loi limite de $\sqrt{n}(\widehat{F}_n(1/2) - 1/4)/\widehat{F}_n(3/4)$ quand $n \to \infty$., où $\widehat{F}_n$ est la fonction de répartition empirique.**

****
****

Exercice 1.15 (Comportement asymptotique des quantiles empiriques)
Si X est une variable aléatoire réelle et $\alpha \in (0, 1)$. Le quantile de X d'ordre α est défini par

$$Q_X(\alpha) = \inf\big(x \in \mathbb{R} : \mathbb{P}[X \leq x] \geq \alpha\big) \tag{1}$$

et la fonction $Q_X : (0, 1) \mapsto \mathbb{R}$ est appelée fonction quantile.

1) Montrer que le quantile d'ordre α de X vérifie

$$\mathbb{P}[X \leq Q_X(\alpha)] \geq \alpha. \tag{2}$$

2) Soit X est une variable aléatoire réelle admettant une densité f_X par rapport à la mesure de Lebesgue et portée par un intervalle I de $\mathbb{R}$. On suppose que f_X est strictement positive sur I presque surement (et nulle en dehors de I). Montrer que la fonction de répartition F_X de X est inversible sur I, sa fonction reciproque F_X^{-1} est continue sur $(0,1)$ et $Q_X(\alpha) = F_X^{-1}(\alpha)$ pour tout $\alpha \in (0,1)$.

On considère l'hypothèse suivante :

Hypothèse (H1) : Dans tout le texte, on suppose que X est une variable aléatoire réelle admettant une densité f_X par rapport à la mesure de Lebesgue. On suppose que f_X est strictement positive (presque surement) sur un intervalle I de $\mathbb{R}$ et nulle en dehors de cet intervalle.

Soit un échantillon $X_1, \ldots, X_n$ de données indépendantes et identiquement distribuées (i.i.d.) de même loi que X. On se fixe $\alpha \in (0,1)$ et on souhaite utiliser les données $X_1, \ldots, X_n$ pour le calcul de $Q_X(\alpha)$. On introduit la fonction de répartition empirique F_n et la fonction quantile empirique Q_n données pour tout $x \in \mathbb{R}$ et $\alpha \in (0,1)$ par

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{X_i \leq x} \text{ et } Q_n(\alpha) = \inf\big(x \in \mathbb{R} : F_n(x) \geq \alpha\big),$$

où, pour tout $1 \leq i \leq n$, $\mathbb{1}_{X_i \leq x}$ vaut 1 quand $X_i \leq x$ et 0 autrement.

3) Soit

$$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)} \tag{3}$$

la suite des réarrangements de l'échantillon. Montrer que sous l'hypothèse (H1), toutes les inégalités de (3) sont strictes presque surement.

4) Soit $\alpha \in (0,1)$. Montrer que le quantile empirique d'ordre α est tel que $Q_n(\alpha) = X_{(\lceil n\alpha \rceil)}$ où $\lceil n\alpha \rceil$ est le plus petit entier supérieur ou égal à $n\alpha$.

5) Montrer que pour tout $\alpha \in (0,1)$, on a presque surement que

$$Q_n(\alpha) \longrightarrow Q_X(\alpha).$$

6) On supposse maintenant que la densité f_X de X admet une version continue en $Q_X(\alpha)$ alors le quantile empirique d'ordre α est asymptotiquement Gaussien :

$$\frac{\sqrt{n} f_X(Q_X(\alpha))}{\sqrt{\alpha(1-\alpha)}}\big(Q_n(\alpha) - Q_X(\alpha)\big) \xrightarrow{\mathcal{L}} \mathcal{N}(0,1).$$

****

2 Vraisemblance, EMV, IC, Information de Fisher

****

Exercice 2.1 (Modèle probit)

Nous disposons d'une information relative au comportement de remboursement ou de

non-remboursement d'emprunteurs :

$$Y_i = \begin{cases} 1 & \textbf{si l'emprunteur } i \textbf{ rembourse}, \\ 0 & \textbf{si l'emprunteur } i \textbf{ est défaillant}. \end{cases}$$

Afin de modéliser ce phénomène, on suppose l'existence d'une variable aléatoire Y_i^* normale, d'espérance m et de variance σ^2, que l'on appellera « capacité de remboursement de l'individu i », telle que :

$$Y_i = \begin{cases} 1 & \textbf{si } Y_i^* > 0, \\ 0 & \textbf{si } Y_i^* \leq 0. \end{cases}$$

On note Φ la fonction de répartition de la loi normale $\mathcal{N}(0,1)$.

1. **Exprimer la loi de Y_i en fonction de Φ.**
2. **Les paramètres m et σ^2 sont-ils identifiables ?**

****
****

Exercice 2.2 (Répartition de génotypes dans une population)

Quand les fréquences de gènes sont en équilibre, les génotypes AA, Aa et aa se manifestent dans une population avec probabilités $(1-\theta)^2$, $2\theta(1-\theta)$ et θ^2 respectivement, où θ est un paramètre inconnu. Plato *et al.* (1964) ont publié les données suivantes sur le type de haptoglobine dans un échantillon de 190 personnes :

Type de haptoglobine	**Hp-AA**	**Hp-Aa**	**Hp-aa**
effectifs	10	68	112

1. **Comment interpréter le paramètre θ ? Proposez un modèle statistique pour ce problème.**
2. **Calculez l'estimateur du maximum de vraisemblance $\hat{\theta}_n$ de θ.**
3. **Donnez la loi asymptotique de $\sqrt{n}(\hat{\theta}_n - \theta)$.**
4. **Proposez un intervalle de confiance de niveau asymptotique 95% pour θ.**

****
****

Exercice 2.3 (Modèle d'autorégression)

On considère les observations $X_1, \ldots, X_n$, où les X_i sont issus du *modèle d'autorégression d'ordred* 1 :

$$X_i = \theta X_{i-1} + \xi_i, \quad i = 1, \ldots, n, \qquad X_0 = 0, \tag{4}$$

où ξ_i i.i.d. de loi normale $\mathcal{N}(0, \sigma^2)$ et $\theta \in \mathbb{R}$.

1. **Explicitez l'expérience statistique associée à la donnée $(X_1, \ldots, X_n)$.**
2. **Calculez l'estimateur du maximum de vraisemblance $\hat{\theta}_n$ de θ pour ce modèle.**

****
****

Exercice 2.4 (Durées de connection)

On peut modéliser la durée d'une connection sur le site `www.Cpascher.com` par une loi gamma$(2, 1/\theta)$ de densité

$$\theta^{-2} x e^{-x/\theta} 1_{[0,+\infty[}(x).$$

Pour fixer vos tarifs publicitaires, vous voulez estimer le paramètre θ à partir d'un échantillon $X_1, \ldots, X_n$ de n durées de connexion. On vous donne $\mathbb{E}_\theta(X_i) = 2\theta$ et $\mathrm{var}_\theta(X_i) = 2\theta^2$.

1. **Calculez l'estimateur du maximum de vraisemblance $\hat{\theta}_n$ de θ.**
2. **Que vaut $\mathbb{E}(\hat{\theta}_n)$? Quelle est la variance de $\hat{\theta}_n$?**

****
****

Exercice 2.5 (Analyse d'un canal de communication)

Une variable aléatoire réelle X suit une loi de Pareto(α, θ) avec $\alpha > 1$ et $\theta > 0$, si elle a pour densité par rapport à la mesure de Lebesgue

$$f_X(x) = \frac{\alpha - 1}{\theta} \left(\frac{\theta}{x} \right)^\alpha \mathbf{1}_{[\theta,\infty[}(x).$$

Les paquets d'information arrivent aléatoirement dans un canal de communication et le temps entre deux paquets est modélisé par une loi de Pareto. On dispose d'un échantillon $X_1, \ldots, X_n$ de temps d'attente, supposés indépendants.

1. **Comment interpréter α et θ ? Vérifier que pour $\alpha > 3$**

$$\mathbb{E}(X) = \frac{\alpha - 1}{\alpha - 2}\theta \quad \textbf{et} \quad \mathbf{Var}(X) = \frac{\alpha - 1}{(\alpha - 3)(\alpha - 2)^2}\theta^2.$$

2. **On suppose $\alpha > 3$ connu. Calculez l'estimateur $\hat{\theta}$ du maximum de vraisemblance de θ.**
3. **Calculez $P(\hat{\theta} > x)$. Quelle est la loi de $\hat{\theta}$? sa moyenne ? sa variance ?**
4. **Que dire du comportement de $\hat{\theta}$ lorsque $n \to \infty$?**
5. **Maintenant, on suppose θ connu, mais pas α. Quel est l'estimateur $\hat{\alpha}$ du maximum de vraisemblance de α ?**
6. **Quelle est la loi de $\log(X_i/\theta)$? de $\sum_{i=1}^n \log(X_i/\theta)$?**
7. **Calculez le biais $\mathbb{E}(\hat{\alpha}) - \alpha$ de $\hat{\alpha}$? Proposez un estimateur non biaisé de α.**
8. **Quelle est la variance de ce dernier estimateur ? Quel est son comportement lorsque $n \to \infty$?**

****
****

Exercice 2.6 (Modèle exponentiel)

Une grande partie des modèles utilisés en pratique sont des modèles exponentiels (modèle gaussien, log-normal, exponentiel, gamma, Bernouilli, Poisson, etc). Nous allons étudier quelques propriétés de ces modèles. On appelle modèle exponentiel une famille de lois $\{\mathbb{P}_\theta,\ \theta \in \Theta\}$ ayant une densité par rapport à une mesure μ σ-finie sur $\mathbb{R}$ ou $\mathbb{N}$ de la forme

$$p_\theta(x) = c(\theta) \exp\left(m(\theta) f(x) + h(x)\right).$$

On supposera que Θ est un ouvert de $\mathbb{R}$, $m(\theta) = \theta$ et $c(\cdot) \in \mathcal{C}^2$, $c(\theta) > 0$ pour tout $\theta \in \Theta$. On notera X une variable aléatoire de loi $\mathbb{P}_\theta$ et on admettra que

$$\frac{\partial^i}{\partial \theta^i} \int \exp(\theta f(x) + h(x))\, \mu(dx) = \int f(x)^i \exp(\theta f(x) + h(x))\, \mu(dx) < +\infty, \quad \textbf{pour } i = 1, 2.$$

1. **Montrez que $\varphi(\theta) := \mathbb{E}_\theta\left(f(X)\right) = -\frac{d}{d\theta}\log(c(\theta))$.**
2. **Montrez que $\mathrm{Var}_\theta(f(X)) = \varphi'(\theta) = -\frac{d^2}{d\theta^2}\log(c\,(\theta))$.**
3. **On dispose d'un n-échantillon $X_1, \ldots, X_n$ de loi $\mathbb{P}_\theta$. On note $\hat{\theta}_n$ l'estimateur obtenu en résolvant $\varphi(\hat{\theta}_n) = \frac{1}{n}\sum_{i=1}^n f(X_i)$. En supposant $\mathrm{Var}_\theta(f(X)) > 0$, montrez que**

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{1}{\mathrm{Var}_\theta\left(f(X)\right)}\right).$$

****
****

Exercice 2.7 (Taux de défaillance)

Une chaîne de production doit garantir une qualité minimale de ses produits. En particulier, elle doit garantir que la proportion θ des produits défaillants reste inférieure à un taux fixé par le client. Un échantillon de n produits est prélevé et analysé. On note $\hat{\theta}_n$ la proportion de produits défectueux dans l'échantillon.

1. **Proposer un modèle statistique pour ce problème. Quelle est la loi de $n\hat{\theta}_n$?**
2. **Quelle information donne la loi des grand nombres et le théorème centrale limite sur le comportement asymptotique de $\hat{\theta}_n$?**
3. **On donne $\mathbb{P}(N > 1.64) = 5\%$ pour $N \sim \mathcal{N}(0,1)$. En déduire ϵ_n (dépendant de n et θ) tel que $\mathbb{P}(\theta \geq \hat{\theta}_n + \epsilon_n) \overset{n\to\infty}{\to} 5\%$.**
4. **La valeur ϵ_n précédente dépend de θ. A l'aide du lemme de Slutsky, donner ϵ'_n ne dépendant que de n et $\hat{\theta}_n$ tel que $\mathbb{P}(\theta \geq \hat{\theta}_n + \epsilon'_n) \overset{n\to\infty}{\to} 5\%$.**

****
****

Exercice 2.8 (Cas des défaillances rares)

La chaîne produit des composants électroniques utilisés dans le secteur aéronautique. Le taux de défaillance doit donc être très bas. En particulier, comme la taille de l'échantillon

ne peut être très grosse (question de coût), il est attendu que θ soir du même ordre de grandeur que $1/n$. On supposera donc par la suite que la proportion de composants défectueux est $\theta_n = \lambda/n$ pour un certain $\lambda > 0$ et on cherche à estimer λ par $\hat{\lambda}_n = n\hat{\theta}_n$. La valeur λ est supposée indépendante de n (le cas intéressant est quand λ est petit).

1. **Quelle est la limite de $\mathbb{P}(\hat{\lambda}_n = k)$ lorsque $n \to +\infty$? En déduire que $\hat{\lambda}_n$ converge en loi vers une variable de Poisson de paramètre λ.**
2. **On suppose qu'il y a une proportion $\theta_n = 3/n$ de composants défectueux. Sachant que $\mathbb{P}(Z = 0) \approx 5\%$ pour Z de loi de Poisson de paramètre 3, montrer que $\mathbb{P}(\theta_n > \hat{\theta}_n + 2/n) \approx 5\%$ pour n grand.**

****
****

Exercice 2.9 (Borne de Cramer-Rao)

On considère un vecteur aléatoire $(X_1, \dots, X_n) \in \mathbb{R}^n$ de loi appartenant à une famille $\{\mathbb{P}_\theta : \theta \in \Theta\}$ de lois sur $\mathbb{R}^n$, avec Θ intervalle ouvert de $\mathbb{R}$. On suppose que $d\mathbb{P}_\theta(x) = p(\theta, x)\, d\mu(x)$ avec μ mesure σ-finie sur $\mathbb{R}^n$ et on note $l_X(\theta) = \log(p(\theta, X))$. On suppose que la famille de lois $\{\mathbb{P}_\theta : \theta \in \Theta\}$ est régulière ; l'information de Fisher $I_n(\theta)$ est donc bien définie et on pourra intervertir intégrales et dérivations à notre de guise.

Pour un estimateur $\hat{\theta}$ donné, on note $R_\theta(\hat{\theta}) := \mathbb{E}_\theta\big(\hat{\theta} - \theta\big)^2$ son risque quadratique et $b(\theta) := \mathbb{E}_\theta[\hat{\theta}] - \theta$ son biais (qu'on suppose dérivable).

1. **Montrez que $R_\theta(\hat{\theta}) = b(\theta)^2 + \mathbf{Var}_\theta(\hat{\theta})$.**
2. **Montrez que $\mathbb{E}_\theta[l'_X(\theta)] = 0$.**
3. **Montrez que $b'(\theta) = \mathbb{E}_\theta\big[\hat{\theta} l'_X(\theta)\big] - 1$.**
4. **Déduire des deux questions précédentes l'égalité $1 + b'(\theta) = \mathbb{E}_\theta\big[\big(\hat{\theta} - \mathbb{E}_\theta(\hat{\theta})\big) l'_X(\theta)\big]$.**
5. **En déduire la borne de Cramér-Rao :**

$$\boxed{R_\theta(\hat{\theta}) \geq \frac{(1 + b'(\theta))^2}{I_n(\theta)} + b(\theta)^2.}$$

6. **Quel est le risque quadratique minimal d'un estimateur sans biais ?**

****
****

Exercice 2.10 (Information de Fisher : entraînement)

Dans les modèles suivants, calculer l'information de Fisher associée aux n observations (si elle est bien définie), l'estimateur du maximum de vraisemblance et sa loi asymptotique :

1. $X_1, \dots, X_n \overset{i.i.d}{\sim} \mathcal{B}(\theta)$.
2. $X_1, \dots, X_n \overset{i.i.d}{\sim} \mathcal{N}(m, v)$.
3. $X_1, \dots, X_n \overset{i.i.d}{\sim} \mathcal{U}[0, \theta]$.

****
****

Exercice 2.11 (Réduction de variance)

L'estimateur du maximum de vraisemblance $\hat{\theta}_{MV}$ est parfois difficile à calculer numériquement et on préfère calculer un estimateur plus simple $\hat{\theta}$ (par exemple l'estimateur des moindres carrés en régression non-linéaire). On va voir qu'on peut en général améliorer simplement $\hat{\theta}$ de façon à avoir un estimateur $\tilde{\theta}$ qui se comporte asymptotiquement comme $\hat{\theta}_{MV}$.

Pour simplifier, on va se placer dans le cadre simple de données i.i.d. Considérons un modèle $(\mathbb{P}_\theta)_{\theta\in\mathbb{R}^d}$ régulier de la forme

$$d\mathbb{P}_\theta(x) = p_\theta(x)\, d\mu(x), \text{ pour tout } x \in \mathbb{R}$$

avec μ mesure de Lebesgue sur $\mathbb{R}$ ou mesure de comptage sur $\mathbb{N}$. On associe à n observations i.i.d. $X = (X_1, \ldots, X_n)$ la log-vraisemblance

$$l_X(\theta) = \sum_{i=1}^{n} \log(p_\theta(X_i))$$

que l'on supposera 3 fois différentiable et l'estimateur du maximum de vraisemblance $\hat{\theta}_{MV}$.

Considérons un autre estimateur $\hat{\theta}$ de θ vérifiant

$$\sqrt{n}(\hat{\theta} - \theta) \overset{loi}{\to} Z \sim \mathcal{N}(0, \Sigma(\theta)).$$

1. **Que dire de $\Sigma(\theta)$ par rapport à l'information de Fisher $I_1(\theta)$ associée à une observation ?**
2. **On note $H_X(\theta)$ la matrice Hessienne associée à la fonction l_X. Montrez que**

$$\hat{\theta}_{MV} = \hat{\theta} - H_X(\hat{\theta})^{-1}\nabla l_X(\hat{\theta}) + O\big((\hat{\theta}_{MV} - \hat{\theta})^2\big).$$

3. **Proposez un estimateur $\tilde{\theta}$ vérifiant $\sqrt{n}(\tilde{\theta} - \hat{\theta}_{MV}) \overset{\mathbb{P}}{\to} 0$.**
4. **Quelle est la loi asymptotique de $\sqrt{n}(\tilde{\theta} - \theta)$? Conclusion ?**

****

3 Tests

****

Exercice 3.1 (Test de Neyman-Pearson)

Chercher la région de rejet du test de Neyman-Pearson dans les cas suivants.

1. **Loi exponentielle $\mathcal{E}(\theta)$. Test de $\theta = \theta_0$ contre $\theta = \theta_1$ avec $\theta_1 > \theta_0$.**
2. **Loi de Bernoulli $\mathcal{B}(\theta)$. Test de $\theta = \theta_0$ contre $\theta = \theta_1$ pour $\theta_1 > \theta_0$. Quel problème rencontre-t-on dans ce cas ?**

****
****

Exercice 3.2 (Test de Wald)

Lors des essais d'un type d'appareils ménagers, une association de consommateurs envisage les 3 issues suivantes : fonctionnement normal, mauvais fonctionnement et défaillance. Les probabilités de fonctionnement normal et de défaillance sont égales à p^2 et à $(1-p)^2$ respectivement, où $p \in]0,1[$ est un paramètre inconnu. Pour un échantillon de $n = 200$ appareils, on a observé que 112 appareils fonctionnent normalement, 12 sont défaillants et 76 fonctionnent mal. A partir de ces données, on cherche à inférer le paramètre p.

1. **Proposer un modèle statistique pour ce problème.**
2. **Chercher l'estimateur du maximum de vraisemblance $\hat{p}_n$ de p. Montrer qu'il est consistant et donner la loi limite de $\sqrt{n}(\hat{p}_n - p)$ quand $n \to \infty$.**
3. **À l'aide du test de Wald, tester l'hypothèse que $p = 1/2$ contre l'alternative $p \neq 1/2$ (on donnera la forme de la région critique et la *p-value* du test). On suppose connues les valeurs de la fonction de répartition de la loi normale standard.**

****
****

Exercice 3.3 (Cancer et tabac)

Voici les chiffres (fictifs) du suivi d'une population de 100 personnes (50 fumeurs, 50 non-fumeurs) pendant 20 ans.

	fumeur	**non-fumeur**
cancer diagnostiqué	**11**	**5**
pas de cancer	**39**	**45**

On s'interroge : la différence du nombre de cancers entre fumeurs et non-fumeurs est-elle statistiquement significative ? On note X_i la variable qui vaut 1 si le fumeur i a été atteint d'un cancer et 0 sinon. De même, on note Y_i la variable qui vaut 1 si le non-fumeur i a été atteint d'un cancer et 0 sinon. On suppose que les X_i sont i.i.d. de loi be Bernoulli $\mathcal{B}(\theta_f)$, les Y_i sont i.i.d. de loi $\mathcal{B}(\theta_{nf})$ et les X_i sont indépendants des Y_i.

1. **Si $\theta_f \neq \theta_{nf}$, quelle est la limite de $\sqrt{n}|\bar{X}_n - \bar{Y}_n|$?**
2. **On suppose que $\theta_f = \theta_{nf} = \theta$ et on note $\hat{\theta} = (\bar{X}_n + \bar{Y}_n)/2$. Montrez que**

$$\sqrt{\frac{n}{2\hat{\theta}(1-\hat{\theta})}}(\bar{X}_n - \bar{Y}_n) \overset{loi}{\to} \mathcal{N}(0,1).$$

3. **Proposez un test de niveau asymptotique 5% de H_0 : "le taux de cancer n'est pas différent" ($\theta_f = \theta_{nf}$) contre H_1 : "le taux de cancer est différent" ($\theta_f \neq \theta_{nf}$).**
4. **Supposons maintenant qu'une étude supplémentaire permet d'avoir le suivi de 300 personnes et que les proportions sont les mêmes :**

	fumeur	**non-fumeur**
cancer diagnostiqué	**33**	**15**
pas de cancer	**117**	**135**

Quelle est la conclusion du test avec ces données ?

5. **Revenons aux chiffres de la première étude : proposez un test de niveau asymptotique 5% de H_0 : "fumer n'a pas d'impact sur le taux de cancer" ($\theta_f = \theta_{nf}$) contre H_1 : "fumer augmente le taux de cancer" ($\theta_f > \theta_{nf}$) ? Quelle est sa conclusion ? Quelle est la p-value associée aux observations ?**
6. **Que retenir ?**

****
****

Exercice 3.4 (Test pour une certification bio)

Pour avoir la certification "bio", un fabriquant de produits "bio" doit garantir pour chaque lot un pourcentage d'OGM inférieur à 1%. Il prélève donc $n = 25$ produits par lot et teste si le pourcentage d'OGM est inférieur à 1%. On note X_i le logarithme naturel du nombre de pourcents d'OGM du paquet numéro i.
Modèle : On suppose que les X_i sont indépendants et suivent une loi gaussienne $\mathcal{N}(\theta, 1)$.

1. **Pour $\theta_1 > \theta_0$, montrez que le test de Neyman-Pearson de niveau α de $H_0 : \theta = \theta_0$ contre $H_1 : \theta = \theta_1$ est de la forme $\bar{X}_n > t_{n,\alpha}$.**
2. **Pour le fabriquant, le pourcentage d'OGM est inférieur à 1% sauf preuve du contraire. Il veut tester l'hypothèse $H_0 : \theta \leq 0$ contre $H_1 : \theta > 0$ et il souhaite que pour $\theta \leq 0$ le test se trompe avec une probabilité inférieure à 5%. Calculez un seuil $t_{25,5}$ tel que**

$$\sup_{\theta \leq 0} \mathbb{P}_\theta(\bar{X}_{25} > t_{25,5}) = 5\%.$$

   On pourra utiliser que $\mathbb{P}(Z > 1.645) \approx 5\%$, pour $Z \sim \mathcal{N}(0, 1)$.
3. **Une association "anti-OGM" veut s'assurer qu'il n'y a effectivement pas plus de 1% d'OGM dans les produits labélisés "bio". En particulier, elle s'inquiète de savoir si le test parvient à éliminer les produits pour lesquels le pourcentage d'OGM dépasse de 50% le maximum autorisé. Quelle est la probabilité que le test ne rejette pas H_0 lorsque le pourcentage d'OGM est de 1.5% ?**
4. **Scandalisée par le résultat précédent, l'association milite pour que le test du fabriquant prouve effectivement que le pourcentage d'OGM est inférieur à 1%. Pour elle, le pourcentage d'OGM est supérieur à 1% sauf preuve du contraire, donc H_0 est $\theta > 0$ et H_1 est $\theta \leq 0$. Proposez un test de H_0 contre H_1 tel que la probabilité que le test rejette à tort H_0 soit inférieure à 5%.**

****
****

Exercice 3.5 (Test de support)

Soient $X_1, \ldots, X_n$ de loi $\mathcal{U}[0,\theta]$ et $M = \max(X_i),\ 1 \leq i \leq n$. On cherche à tester $H_0 : \theta = 1$ contre $H_1 : \theta > 1$.

1. **Pourquoi ne peut-on pas utiliser ici le test de Neyman-Pearson ?**
2. **On propose le test suivant : on rejette H_0 lorsque $M > c$ (c constante donnée). Calculer la fonction de puissance.**
3. **Quelle valeur prendre pour c pour obtenir un niveau de 5% ?**
4. **Si $n = 20$ et que la valeur observée de M est 0.96, que vaut la p-value ? quelle conclusion tirer sur H_0 ? Même question pour $M^{obs} = 1.04$.**

****
****

Exercice 3.6 (Peut-on retarder sa mort ?)

On prétend couramment que les mourants peuvent retarder leur décès jusqu'à certains événements importants. Pour tester cette théorie, Philips et King (1988, article paru dans *The Lancet*, prestigieux journal médical) ont collecté des données de décès aux environs d'une fête religieuse juive. Sur 1919 décès, 922 (resp. 997) ont eu lieu la semaine précédente (resp. suivante). Comment utiliser de telles données pour tester cette théorie grâce à un test asymptotique ?

****
****

Exercice 3.7 (Tests multiples pour puces à ADN)

Après un traitement approprié, les données de puces à ADN correspondent à un grand vecteur $(Y_1, \ldots, Y_p)$ de différences de log-intensités. Typiquement le nombre p de gènes est de l'ordre de quelques milliers. Ces observations peuvent être modélisées comme suit : $Y_j = \theta_j + \varepsilon_j$, pour $j = 1, \ldots, p$ avec les ε_j i.i.d. de loi $\mathcal{N}(0, \sigma^2)$ avec σ^2 connu. Les sites qui nous intéressent d'un point de vue biologique sont les sites j où θ_j est non nul (on dit alors que le site j est positif). En général, de 1 à 10% des sites sont positifs.

1. **Pour chaque j proposez un test de niveau 5% de $H_0 : \theta_j = 0$ contre $H_1 : \theta_j \neq 0$.**
2. **Supposons que $p = 5000$ et 4% des sites sont positifs. Quel est le nombre moyen de faux positifs que va engendrer la famille de tests ci-dessus ?**
3. **Pour $\gamma > 0$ calculez la limite de $\mathbb{P}(\max_{j=1,\ldots,p} |\varepsilon_j| > \gamma\sqrt{2\sigma^2 \log p})$ lorsque $p \to \infty$.**
   Formule : $\int_T^\infty e^{-x^2/2}\, dx \sim T^{-1} e^{-T^2/2}$ lorsque $T \to \infty$.
4. **Proposez un nouveau seuil pour les tests de $H_0 : \theta_j = 0$ contre $H_1 : \theta_j \neq 0$ tel que la probabilité que les p tests ne donnent aucun faux positif soit minorée par $1 - \alpha$.**
5. **Pour p de l'ordre de quelques milliers, que pensez-vous de la puissance du test ci-dessus ? Sachant que les sites positifs donnent lieu à de nouvelles expériences**

biologiques et que celles-ci coûtent cher, proposez une autre notion du niveau de confiance qui permettrait une meilleur puissance tout en limitant les coûts liés à des faux positifs.

****
****

Exercice 3.8 (Comment dimensionner une enquête ?)

Quelle taille d'échantillon faut-il retenir pour connaître à deux points de pourcentage près (au plus) avec 95 chances sur 100, la proportion de parisiens qui portent des lunettes ? On suppose que chaque parisien a la même probabilité d'être sondé et que les individus sont tirés avec remise.

****
****

Exercice 3.9 (Test dans un modèle ANOVA)

On dispose de l'observation de variables aléatoires

$$Y_{ij} = m_i + \xi_{ij}, \quad i = 1, \dots, k, \quad j = 1, \dots, l,$$

où $(m_1, \dots, m_k) \in \mathbb{R}^k$ **et les** ξ_{ij} **sont des variables aléatoires i.i.d. de loi** $\mathcal{N}(0, \sigma^2)$**. On considère le problème de test d'égalité des moyennes**

$$\mathbf{H}_0 : m_1 = m_2 = \cdots = m_k \quad \textbf{contre} \quad \mathbf{H}_1 : \ \exists i \neq i' \ \textbf{tels que} \ m_i \neq m_{i'}.$$

1. **Montrer qu'il s'agit d'un modèle de régression linéaire avec la matrice** $\mathbf{X}$ **que l'on précisera, ainsi que la matrice** $B = \mathbf{X}^T\mathbf{X}$ **correspondante.**
2. **Montrer que l'hypothèse nulle s'écrit**

$$\mathbf{H}_0 : \ m \in \Theta_0 = \{m : Gm = 0\}$$

   avec la matrice G **que l'on précisera.**
3. **En déduire la forme du test de Fisher dans ce contexte.**

****
****

Exercice 3.10 (Test d'adéquation du χ^2)

Une vaste enquête au sein d'un central téléphonique a permis de déterminer que le nombre d'appels reçus durant une seconde suit une loi de Poisson de paramètre 4.

Pour un second central téléphonique, on effectue une enquête de moindre envergure afin de vérifier si le nombre d'appels reçus par seconde suit la même loi. On comptabilise lors de 200 secondes, le nombre d'appels par seconde, ce qui produit les résultats suivants :

Nombre d'appels par seconde	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
Effectifs observés	**6**	**15**	**40**	**42**	**37**	**30**	**10**	**9**	**5**	**3**	**2**	**1**

On note N_i le nombre d'appels reçus entre les secondes i et $i+1$ et on suppose les N_i indépendantes. Enfin on note $N_i' = \min(N_i, 8)$.

1. **On note Q la loi de $\min(N, 8)$ où N suit une loi de Poisson de paramètre 4. Tester l'adéquation des N_i' observés à cette loi Q.**
2. **On suppose que les temps entre deux appels consécutifs suivent une loi exponentielle de paramètre λ. Dans ce cas, les N_i sont indépendants et suivent une loi de Poisson de paramètre λ. Proposez un test de niveau asymptotique 5% de $\lambda = 4$ contre $\lambda \neq 4$.**
3. **Conclusion ?**

Table de la loi de Poisson :

k	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
$\mathbb{P}_{3.7}[X=k]$	**.025**	**.091**	**.169**	**.209**	**.193**	**.143**	**.088**	**.047**	**.021**	**.009**	**.003**	**.001**	3.10^{-4}	10^{-4}
$\mathbb{P}_4[X=k]$	**.018**	**.073**	**.146**	**.195**	**.195**	**.156**	**.104**	**.059**	**.03**	**.013**	**.005**	**.002**	6.10^{-4}	2.10^{-4}

****

4 Modèle de régression

****

Exercice 4.1 (Modèle de régression simple)

On considère le modèle de regression simple

$$y = ae + bx + \xi,\ \mathbb{E}[\xi] = 0,\ \mathbb{E}[\xi\xi^T] = \sigma^2 I_n,$$

où y, e, x, ξ sont vecteurs de $\mathbb{R}^n$, avec $e = (1, 1, \ldots, 1)^T$ et $x = (x_1, x_2, \ldots, x_n)^T$. Les paramètres réels a et b sont inconnus.

1. **Donner les expressions des estimateurs des moindres carreés (MC) $\hat{a}$ et $\hat{b}$ des paramètres a et b .**
2. **Montrer que la droite de régression définie par l'équation**

$$y^* = \hat{a} + \hat{b}x^* \quad , (x^*, y^*) \in \mathbb{R}^2,$$

   passe par le point moyen $(\bar{x}, \bar{y})$ du nuage de points $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$.
3. **On note $\hat{y}_i = \hat{a} + \hat{b}x_i$. Montrer que $\bar{\hat{y}} = \bar{y}$.**
4. **On définit le *coefficient de détermination* R^2 par**

$$R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}.$$

   Montrer l'équation d'analyse de la variance

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

   et en déduire que $R^2 \in [0, 1]$.

5. **Montrer que le coefficient de détermination est égal au carré du coefficient de corrélation linéaire entre y et x.**
6. **Calculer $\mathbf{Var}(\hat{a})$, $\mathbf{Var}(\hat{b})$ et la matrice de covariance du vecteur $(\hat{a}, \hat{b})$ notée $\mathbf{Var}(\hat{a}, \hat{b})$.**
7. **Donner l'estimateur $\hat{\sigma}^2$ de σ^2 basé sur $\hat{a}, \hat{b}$. En déduire des estimateurs sans biais de $\mathbf{Var}(\hat{a})$, $\mathbf{Var}(\hat{b})$ et $\mathbf{Var}(\hat{a}, \hat{b})$.**
8. **On suppose dorénavant $a = 0$ et donc**

$$y = bx + \xi,\ \mathbb{E}[\xi] = 0,\ \mathbb{E}[\xi\xi^T] = \sigma^2 I_n.$$

   Calculer $\tilde{b}$ (l'estimateur des MC de b) et $\tilde{\sigma}^2$ (l'estimateur de σ^2) dans ce nouveau modèle. L'estimateur $\tilde{b}$ est-il égal à $\hat{b}$?
9. **La droite de régression définie par l'équation**

$$y^* = \tilde{b}x^* \quad , (x^*, y^*) \in \mathbb{R}^2,$$

   passe-t-elle par le point moyen $(\bar{x}, \bar{y})$ du nuage de points $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$? Pourquoi ?
10. **Que dire du R^2 dans ce nouveau modèle ?**

****
****

Exercice 4.2 (Modèle de régression multiple)

On considère le modèle de regression multiple

$$y = \theta_0 e + X\theta + \xi, \quad \textbf{où}\ \ \mathbb{E}[\xi] = 0,\ \mathbb{E}[\xi\xi^T] = \sigma^2 I_n,\ e = (1, 1, \ldots, 1)^T$$

avec X une matrice $n \times k$ de rang k et y, ξ des vecteurs de $\mathbb{R}^n$. Les paramètres $\theta_0 \in \mathbb{R}$ et $\theta \in \mathbb{R}^k$ sont inconnus. On note $\hat{\theta}_0$ et $\hat{\theta}$ les estimateurs des moindres carrés de θ_0 et θ.

1. **On note $\hat{y} = \hat{\theta}_0 e + X\hat{\theta}$. Montrer que $\bar{\hat{y}} = \bar{y}$, où $\bar{y}$ (resp. $\bar{\hat{y}}$) est la moyenne des y_i (resp. des $\hat{y}_i$). En déduire que $\bar{y} = \hat{\theta}_0 + \bar{X}\hat{\theta}$ où $\bar{X} = \frac{1}{n} e^T X = \left(\bar{X}_1, \ldots, \bar{X}_k\right)$.**
2. **Montrer l'équation d'analyse de la variance :**

$$\|y - \bar{y}e\|^2 = \|y - \hat{y}\|^2 + \|\hat{y} - \bar{y}e\|^2.$$

   En déduire que le *coefficient de détermination*

$$R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

   est toujours inférieur à 1.
3. **Supposons que $Z = [e, X]$ est de rang $k + 1$. Calculez en fonction de Z la matrice de covariance de $(\hat{\theta}_0, \hat{\theta})$. Comment accède-t-on à $\mathbf{Var}(\hat{\theta}_j)$, pour $j = 0, \ldots, p$?**

4. **On suppose dorénavant que $\theta_0 = 0$ et donc**

$$y = X\theta + \xi, \quad \mathbb{E}[\xi] = 0,\ \mathbb{E}[\xi\xi^T] = \sigma^2 I_n.$$

L'estimateur des moindres carrés $\tilde{\theta}$ dans ce modèle est-il égal à $\hat{\theta}$?

5. **A-t-on la relation $\bar{\hat{y}} = \bar{y}$? Que dire du R^2 dans ce modèle ?**

****
****

Exercice 4.3 (Régression Ridge)
On considère le modèle de regression

$$\underset{(n,1)}{Y} = \underset{(n,k)}{X}\,\underset{(k,1)}{\theta} + \underset{(n,1)}{\xi}\,.$$

On suppose que X est une matrice déterministe, $\mathbb{E}[\xi] = 0$, $\mathbb{E}[\xi\xi^T] = \sigma^2 I_n$,

1. **On suppose que $k > n$. Que dire de l'estimation par moindres carrés ?**
2. **On appelle estimateur `Ridge regression` de paramètre de régularisation $\lambda > 0$ l'estimateur**

$$\hat{\theta}_\lambda = \arg\min_{\theta\in\mathbb{R}^k} \left\{ \|Y - X\theta\|^2 + \lambda\|\theta\|^2 \right\}.$$

Exprimez $\hat{\theta}_\lambda$ en fonction de X, Y et λ. Cet estimateur est-il défini pour $k > n$?

3. **Calculez la moyenne et la matrice de covariance de l'estimateur Ridge. Est-il sans biais ?**
4. **On suppose maintenant que $k = 1$, ce qui correspond au modèle de régression simple. Montrer qu'il existe une valeur de λ telle que le risque de l'estimateur Ridge de paramètre λ est inférieur au risque de l'estimateur des MC.**

****
****

Exercice 4.4 (Théorème de Gauss-Markov)
On considère le modèle de regression

$$\underset{(n,1)}{Y} = \underset{(n,k)}{X}\,\underset{(k,1)}{\theta} + \underset{(n,1)}{\xi}\,.$$

On suppose que X est une matrice déterministe, $\mathbb{E}[\xi] = 0$, $\mathbb{E}[\xi\xi^T] = \sigma^2 I_n$, Rang$(X) = k$. On note $\hat{\theta}$ l'estimateur des MC de θ.

1. **Montrer que $\hat{\theta}$ est sans biais et expliciter sa matrice de covariance.**
2. **Soit $\tilde{\theta}$ un estimateur de θ linÈaire en Y, i.e., $\tilde{\theta} = LY$ pour une matrice $L \in \mathbb{R}^{k\times n}$ déterministe. Donner une condition nÈcessaire et suffisante sur L pour que $\tilde{\theta}$ soit sans biais. On supposera maintenant cette hypothËse vÈrifiÈe.**

3. **Calculer la matrice de covariance de $\tilde{\theta}$. En posant $\Delta = L - (X^TX)^{-1}X^T$ montrer que $\Delta X = 0$ et $\mathbf{cov}(\tilde{\theta}) = \mathbf{cov}(\hat{\theta}) + \sigma^2\Delta\Delta^T$. En déduire que**

$$\mathbb{E}[(\tilde{\theta}-\theta)(\tilde{\theta}-\theta)^T] \geq \mathbb{E}[(\hat{\theta}-\theta)(\hat{\theta}-\theta)^T] \quad \textbf{(inégalité au sens matriciel).}$$

4. **En passant au risques quadratiques $\mathbb{E}\big[\|\tilde{\theta}-\theta\|^2\big]$ et $\mathbb{E}\big[\|\hat{\theta}-\theta\|^2\big]$, en déduire que l'estimateur des MC est optimal dans la classe de tous les estimateurs linéaires sans biais.**

****
****

Exercice 4.5 (Phénomène de Stein)

On considère le modèle

$$Y_j = \theta_j + \xi_j,\ j = 1,\dots,d,$$

avec les ξ_j iid gaussiennes centrées de variance 1. On pose $Y = (Y_1,\dots,Y_d)$ et $\theta = (\theta_1,\dots,\theta_d)$. On s'intéresse à l'estimation de θ et on suppose $d \geq 3$.

Definition : Un estimateur θ^* de θ est dit admissible sur $\Theta \subset \mathbb{R}^d$ par rapport au risque quadratique s'il n'existe pas d'estimateur $\hat{\theta}$ tel que pour tout $\theta \in \Theta$

$$\mathbb{E}_\theta\Big[\|\hat{\theta}-\theta\|^2\Big] \leq \mathbb{E}_\theta\Big[\|\theta^*-\theta\|^2\Big],$$

avec inégalité stricte en au moins un $\theta_0 \in \Theta$.
Lemme (admis) Soit $d \geq 3$. Pour tout $\theta \in \mathbb{R}^d$, on a $0 < \mathbb{E}[\|Y\|^{-2}] < \infty$.

1. **Donner l'estimateur intuitif de θ. Quel est son biais ? sa variance ? son risque quadratique ?**
2. **Soit $\xi = (\xi_1,\dots,\xi_d)$. Montrer que si $f : \mathbb{R}^d \to \mathbb{R}$ vérifie**
   (a) **pour presque tout $(x_2,\dots,x_d)$, la fonction $x_1 \to f(x_1,\dots,x_d)$ est dérivable et $\lim_{|x_1|\to\infty} f(x_1,\dots,x_d)e^{-x_1^2/2} = 0$,**
   (b) $\mathbb{E}[|\frac{\partial f}{\partial x_1}(\xi)|] < +\infty$,
   alors $\mathbb{E}[\frac{\partial f}{\partial x_1}(\xi)] = \mathbb{E}[\xi_1 f(\xi)]$. $\quad (E)$
3. **Soit $\xi = (\xi_1,\dots,\xi_d)$. Soit $f : \mathbb{R}^d \to \mathbb{R}$ telle que $f(u_1,\dots,u_d)$ admet des dérivées partielles par rapport à chaque composantes pour presque toutes les valeurs des autres composantes $\mathbb{E}[|\frac{\partial f}{\partial x_i}(\xi)|] < +\infty,\ i = 1,\dots,d$.**
   On veut montrer que

$$\mathbb{E}[\xi_1 f(\xi)] = \mathbb{E}[\frac{\partial f}{\partial x_1}(\xi)], \quad \textbf{(E)}$$

   (ce qui implique que pour tout $i = 1,\dots,d$, $\mathbb{E}[\xi_i f(\xi)] = \mathbb{E}[\frac{\partial f}{\partial x_i}(\xi)]$. On note (E') l'égalité

$$\int_{-\infty}^{+\infty} u f(u,x_2,\dots,x_d)e^{-u^2/2}du = \int_{-\infty}^{+\infty} \frac{\partial f}{\partial u}(u,x_2,\dots,x_d)e^{-u^2/2}du.$$

   Montrer que (E') implique (E).

4. **Montrer que**
$$e^{-u^2/2} = I_{u\geq 0}\int_u^{+\infty} ze^{-z^2/2}dz + I_{u<0}\int_{-\infty}^{u} ze^{-z^2/2}dz.$$

5. **Montrer l'égalité (E').**

6. **Montrer que si $\tilde{f}:\mathbb{R}^d\to\mathbb{R}$ vérifie**
   (a) **$\tilde{f}(y_1,\ldots,y_d)$ admet des dérivées partielles par rapport à chaque composantes pour presque toutes les valeurs des autres composantes.**
   (b) **$\lim_{|y_i|\to\infty}\tilde{f}(y_1,\ldots,y_d)e^{-(y_i-\theta_i)^2/2}=0$ pour presque tout $(y_1,\ldots,y_{i-1},y_{i+1},\ldots,y_n)$, et tout $i=1,\ldots,d$,**
   (c) **$\mathbb{E}[|\frac{\partial\tilde{f}}{\partial y_i}(Y)|]<+\infty,\ i=1,\ldots,d$,**

   alors
$$\mathbb{E}[(Y_i-\theta_i)\tilde{f}(Y)] = \mathbb{E}[\frac{\partial\tilde{f}}{\partial y_i}(Y)],\ i=1,\ldots,d.$$

7. **Soit $g:\mathbb{R}^d\to\mathbb{R}$ telle que les conditions de la question prÈcÈdente sont vÈrifiÈes par les $f_i(y)=(1-g(y))y_i,\ i=1,\ldots,d$. On considère un estimateur de la forme $\hat{\theta}=g(Y)Y$ (i.e. $\hat{\theta}_j=g(Y)Y_j$). Montrer que**
$$\mathbb{E}[\|\hat{\theta}-\theta\|^2] = d+\mathbb{E}[W(Y)],$$

   avec
$$W(Y) = -2d(1-g(Y)) + 2\sum_{i=1}^{d} Y_i\frac{\partial g}{\partial y_i}(Y) + \|Y\|^2(1-g(Y))^2.$$

8. **Soit $g(y)$ de la forme $1-\frac{c}{\|y\|^2}$. Dans ce cas les $f_i(y)=(1-g(y))y_i$ vérifient les hypothèses de la question 4. Trouver c tel que $\mathbb{E}[W(Y)]<0$.**

9. **L'estimateur intuitif est-il admissible ?**

****
****

Exercice 4.6 (Théorème de Frisch-Waugh)
On considère le modèle linéaire suivant

$$\underset{(T,1)}{Y} = \underset{(T,k_1)(k_1,1)}{X_1\ b_1} + \underset{(T,k_2)(k_2,1)}{X_2\ b_2} + \underset{(T,1)}{u}, \tag{5}$$

On suppose que X_1 et sont X_2 déterministes, $\mathbb{E}[u]=0$, $\mathbf{Var}(u)=\sigma^2 I_T$, $\mathbf{Rang}(X_1)=k_1$, $\mathbf{Rang}(X_2)=k_2$.

1. **Montrer que**
$$||Y-X_1b_1-X_2b_2||^2 = ||M_1Y-M_1X_2b_2||^2 + ||P_1Y-X_1b_1-P_1X_2b_2||^2,$$

   où $P_1=X_1(X_1'X_1)^{-1}X_1$ est le projecteur orthogonal sur l'espace vectoriel engendré par les vecteurs colonnes de X_1 et $M_1=I-P_1$.

2. **En déduire que l'estimateur $\hat{b}_2$ de b_2 obtenu en appliquant la méthode des moindres carrés ordinaires au modèle (5) coïncide avec l'estimateur $\tilde{b}_2$ de b_2 obtenu en appliquant la méthode des moindres carrés ordinaires au modèle suivant**

$$M_1 Y = M_1 X_2 b_2 + v, \ ,\mathbb{E}[v] = 0, \ \mathbf{Var}(v) = \sigma^2 I_T. \tag{6}$$

3. **Montrer que $\tilde{b}_2$ admet les trois écritures suivantes :**

$$\tilde{b}_2 = (\tilde{X}_2'\tilde{X}_2)^{-1}\tilde{X}_2'\tilde{Y} = (\tilde{X}_2'\tilde{X}_2)^{-1}X_2'\tilde{Y} = (\tilde{X}_2'\tilde{X}_2)^{-1}\tilde{X}_2'Y,$$

avec $\tilde{Y} = M_1 Y$ et $\tilde{X}_2 = M_1 X_2$.

4. **En déduire que $\hat{b}_2$ peut être obtenu comme l'estimateur de b_2 dans le modèle :**

$$Y = M_1 X_2 b_2 + \omega.$$

5. **En déduire la valeur de $\hat{b}_1$ en fonction de $\hat{b}_2$. Interpréter les formules obtenues.**
6. **Si on suppose $X_2'X_1 = 0$, déduire de ce qui précède que les estimateurs des mco $\hat{b}_1$ et $\hat{b}_2$ de b_1 et b_1 peuvent être obtenus séparèment en appliquant les mco aux modèles suivants :**

$$Y = X_1 b_1 + u \text{ et } Y = X_2 b_2 + u.$$

****
****

Exercice 4.7 (EMC)
Soient $\xi_1, \ldots, \xi_n$ des variables aléatoires i.i.d. de densité $f(\cdot)$ par rapport à la mesure de Lebesgue sur $\mathbb{R}$, et soit $X_i \in \mathbb{R}, \ i = 1, \ldots, n$. On observe les couples $(X_i, Y_i), \ i = 1, \ldots, n$, issus du modèle de régression linéaire

$$Y_i = \theta X_i + \xi_i, \quad i = 1, 2, \ldots,$$

où $\theta \in \mathbb{R}$ est un paramètre inconnu.

1. **On suppose d'abord que $X_i = i$ et la loi de ξ_i est $\mathcal{N}(0,1)$. $\sigma^2 = 1$. les X_i sont déterministes.**
   (a) **Notons $\hat{\theta}$ l'estimateur des moindres carrés de θ. Quelle est la loi de $\hat{\theta}$?**
   (b) **Expliciter la densité jointe de $Y_1, \ldots, Y_n$.**
   (c) **Supposons que la loi de ξ_i est $\mathcal{N}(0, \sigma^2)$. Déduire de 1.1) l'estimateur du maximum de vraisemblance. Quelle est la loi de $\hat{\theta}^{MV}$? Son risque quadratique ?**
2. **On suppose maintenant que les X_i sont des variables aléatoires i.i.d. et que X_i est indépendant de ξ_i pour tout i.**
   (a) **Soit $\hat{\theta}_n$ l'estimateur des moindres carrés de θ. En supposant que les ξ_i sont de moyenne $\mathbb{E}(\xi_1) = 0$ et de variance $\mathbb{E}(\xi_1^2) = \sigma_\xi^2$ et que $\mathbb{E}(X_1^2) = \sigma_X^2$, donner la loi asymptotique de $\sqrt{n}(\hat{\theta}_n - \theta)$ quand $n \to \infty$.**
   (b) **En déduire un intervalle de confiance de niveau asymptotique $1 - \alpha$ pour θ.**

****

5 Statistiques Bayésiennes

****

Exercice 5.1 (Calcul de loi a posteriori dans le modèle de Bernoulli)

Soient $X_1, \ldots, X_n$ n variables aléatoires indépendantes de loi de Bernoulli $\mathcal{B}(\theta)$ et une loi a priori sur $\Theta = [0,1]$ donnée par la loi uniforme sur $[0,1]$, càd $\mathbb{P}^\theta \sim \mathcal{U}([0,1])$. Montrer que la loi a posteriori de $\theta | X_1, \ldots, X_n$ est une loi Beta de paramètre $\alpha = S_n + 1$ et $\beta = n - S_n + 1$ où $S_n = \sum_{i=1}^n X_i$, càd de densité

$$f(\theta | X_1, \ldots, X_n) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1} = \frac{\Gamma(n+2)}{\Gamma(S_n+1)\Gamma(n-S_n+1)} \theta^{S_n}(1-\theta)^{n-S_n}.$$

Maintenant, on suppose que $p \sim Beta(\alpha, \beta)$ (le cas de la loi uniforme est obtenue pour $\alpha = \beta = 1$). Montrer que $\theta | X_1, \ldots, X_n \sim Beta(\alpha + S_N, \beta + n - S_n)$.

***Remarque :* On voit ici que si la mesure a priori est une loi Beta alors la mesure a posteriori est aussi une loi Beta – de paramètres différents. Quand la prior et la posterior appartiennent à la même famille de loi, on dit que la mesure a priori est conjuguée au modèle.**

****

6 Exercices supplémentaires

****

Exercice 6.1 (Réduction de dimension : Analyse en Composantes Principales (ACP))

Reprenons le modèle de regression ci-dessus avec $p > n$. On supposera que les colonnes de la matrice $X = \begin{bmatrix} X_1^T \\ \vdots \\ X_n^T \end{bmatrix}$ sont centrées. Une technique classique pour gérer ce type de situation est de projeter les $X_i \in \mathbb{R}^p$ sur un espace vectoriel V de petite dimension k. Pour ne pas perdre trop d'information, il est nécessaire que les projections des X_i sur V approximent "au mieux" les X_i. On va rechercher des espaces $V_k =$Vect$\{\alpha^{(1)}, \ldots, \alpha^{(k)}\}$ tels que

$$\sum_{i=1}^n \|X_i - \mathbf{Proj}_{V_k}(X_i)\|^2$$

est minimal et les $\alpha^{(1)}, \ldots, \alpha^{(k)} \in \mathbb{R}^p$ sont de norme 1 et orthogonaux entres eux.

1. **On s'intéresse d'abord au cas où $k = 1$. Montrez que chercher $\alpha \in \mathbb{R}^p$ de norme 1 minimisant**

$$\sum_{i=1}^n \|X_i - \mathbf{Proj}_{<\alpha>}(X_i)\|^2$$

revient à chercher $\alpha \in \mathbb{R}^p$ maximisant

$$\max_{\|\alpha\|=1} \alpha^T (X^T X) \alpha.$$

2. **Comment obtient-on $\alpha^{(1)}$?**
3. **Pour $k \leq$rang(X), comment obtient-on $\alpha^{(1)}, \ldots, \alpha^{(k)}$?**
4. **Montrer que les C_j définis par $C_j = X\alpha^{(j)}$ sont orthogonaux entres eux.**
5. **Exprimer en fonction des valeurs propres de $X^T X$ le rapport**

$$\frac{\sum_{i=1}^n \|X_i - \mathbf{Proj}_{V_k}(X_i)\|^2}{\sum_{i=1}^n \|X_i\|^2}\,.$$

****

****

Exercice 6.2 (Amélioration d'estimateur)

Soit $(Y_1, \ldots, Y_n)$ iid de densité $f(t) = \theta e^{-\theta t} I_{t\geq 0}$. Soit F la fonction de répartition des Y_i. On souhaite estimer la fonction de survie en t : $\bar{F}(t) = 1 - F(t)$.

1. **Proposer un estimateur de $\widehat{\bar{F}(t)}_n$ de $\bar{F}(t)$ qui soit sans biais quelle que soit la loi des (Y_i). Intuitivement, cet estimateur est-il de variance minimale parmi les estimateurs sans biais.**
2. **Calculer la loi limite de $\sqrt{n}(\widehat{\bar{F}(t)}_n - F(t))$.**
3. **Soit T défini par $T = I_{Y_1 > t}$. On note $S = Y_1 + \cdots + Y_n$. Déterminer la loi de Y_1 conditionnellement à S.**
4. **Calculer $T^* = E[T|S]$.**
5. **T^* est-il un estimateur ? Si oui calculer son biais.**
6. **Comparer les variance de T et T^*.**

****

****

Exercice 6.3 (Estimation bayésienne et minimax)

On dispose d'une observation $X = (X_1, \ldots, X_n) \in \mathbb{R}^n$ dont la loi appartient à une famille $(\mathbb{P}_\theta)_{\theta\in\mathbb{R}}$ de lois sur $\mathbb{R}^n$. On suppose qu'il existe une mesure positive σ-finie μ (typiquement μ est la mesure de Lebesgue sur $\mathbb{R}^n$ ou la mesure de comptage sur $\mathbb{N}^n$) telle que

$$d\mathbb{P}_\theta(x) = p(\theta, x)\, d\mu(x), \quad \textbf{pour tout } x \in \mathbb{R}^n \textbf{ et } \theta \in \mathbb{R}.$$

A toute fonction mesurable $\hat{\theta} : \mathbb{R}^n \to \mathbb{R}$, on associe l'estimateur $\hat{\theta}(X)$ de θ. Son risque quadratique est $R_\theta(\hat{\theta}) = \mathbb{E}_\theta\left\{\left(\theta - \hat{\theta}(X)\right)^2\right\} \in \mathbb{R}^+ \cup \{+\infty\}$. Lorsqu'on ne dispose d'aucune information *a priori* sur θ, il est naturel de considérer le risque minimax de $\hat{\theta}$:

$$R^*(\hat{\theta}) = \sup_{\theta\in\mathbb{R}} R_\theta(\hat{\theta}).$$

On suppose maintenant que le paramètre inconnu θ_0 de la loi $\mathbb{P}_{\theta_0}$ de X est le résultat d'un tirage selon une loi π sur $\mathbb{R}$ <u>connue</u>, appelée loi *a priori*. On supposera que $\int_{\mathbb{R}} \theta^2\, d\pi(\theta) < +\infty$. Dans ce cas, il est naturel de considérer le risque dit bayésien de $\hat{\theta}$:

$$R^\pi(\hat{\theta}) = \int_{\mathbb{R}} R_\theta(\hat{\theta})\, d\pi(\theta) \;\; \in \mathbb{R}^+ \cup \{+\infty\}\,.$$

1. **Montrez que $R^\pi(\hat\theta) \le R^*(\hat\theta)$ pour tout $\hat\theta : \mathbb{R}^n \to \mathbb{R}$ mesurable.**
2. **On note Q la probabilité sur $\mathbb{R}\times\mathbb{R}^n$ définie par $dQ(\theta,x) = p(\theta,x)\,d\pi(\theta)d\mu(x)$. Montrez que $R^\pi(\hat\theta) = \mathbb{E}\left\{(\tilde\theta - \hat\theta(\tilde X))^2\right\}$ où $(\tilde\theta,\tilde X)$ est distribué selon la loi Q.**
3. **En déduire que $R^\pi(\hat\theta)$ est minimal en**

$$\hat\theta^\pi(x) = \int_{\mathbb{R}} \theta\, d\pi(\theta|x) = \mathbb{E}_Q\left\{\tilde\theta \mid \tilde X = x\right\}$$

$$\text{où} \quad d\pi(\theta|x) = \frac{p(\theta,x)}{\int_{\mathbb{R}} p(\alpha,x)\,d\pi(\alpha)}\, d\pi(\theta) \quad (\text{avec la convention } 0/0 = 0),$$

est appelée loi *a posteriori*. **L'estimateur $\hat\theta^\pi(X)$ est l'estimateur bayésien de θ.**

4. **Pour une application mesurable $\hat\theta : \mathbb{R}^n \to \mathbb{R}$ donnée, on suppose qu'il existe une suite $(\pi_k)_{k\ge0}$ de lois de probabilité sur $\mathbb{R}$ telle que $R^{\pi_k}(\hat\theta^{\pi_k}) \to R^*(\hat\theta)$ lorsque $k\to\infty$. Montrer que $R^*(\hat\theta) = \inf_{\tilde\theta} R^*(\tilde\theta)$, où l'inf est pris sur toutes les fonctions mesurables $\tilde\theta : \mathbb{R}^n \to \mathbb{R}$. Que dire de $\hat\theta(X)$?**

****
****

Exercice 6.4 (Estimateur bayésien)

Le but de cet exercice est d'introduire une classe de méthodes d'estimation dites bayésiennes. Ces méthodes sont souvent utilisées dans la pratique quand on dispose d'une information supplémentaire sur le paramètre à estimer θ de type que certaines valeurs de θ sont *a priori* **"plus probables" que les autres. Cette information est résumée en termes d'une densité de probabilité $\pi(\theta)$ dite** *densité a priori* **sur l'ensemble des paramètres Θ supposée connue au statisticien.**

On dispose des observations $X_1,\dots,X_n$, où les X_i sont i.i.d. de densité $f(x,\theta^*)$ par rapport à la mesure de Lebesgue sur $\mathbb{R}$ appartenant à une famille paramétrique connue $\{f(x,\theta), \theta\in\Theta\}$ de densités sur $\mathbb{R}$, où $\Theta\subseteq\mathbb{R}$ est une ensemble donné.

Soit $\hat\theta = \hat\theta(X_1,\dots,X_n)$ un estimateur de θ et soit

$$R_n(\theta,\hat\theta) = E_\theta[(\hat\theta-\theta)^2]$$

son risque quadratique. Soit $\pi(\cdot)$ une densité de probabilité par rapport à la mesure de Lebesgue sur $\mathbb{R}$ (densité a priori de θ) telle que $\int_\Theta \pi(\theta)d\theta = 1$ et $\int_\Theta \theta^2\pi(\theta)d\theta < \infty$.

Le risque bayésien **de $\hat\theta$ est défini par :**

$$R_n^\pi(\hat\theta) = \int_\Theta R_n(\theta,\hat\theta)\pi(\theta)d\theta.$$

L'estimateur bayésien **de θ noté $\hat\theta^\pi$ est un estimateur qui fournit le minimum du risque bayésien :**

$$R_n^\pi(\hat\theta^\pi) = \min_{\hat\theta} R_n^\pi(\hat\theta),$$

où $\min_{\hat\theta}$ désigne le minimum sur tous les estimateurs.

1. **On note p la densité de probabilité sur $\mathbb{R}^{n+1}$ définie par**

$$p(\theta, x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i, \theta)\pi(\theta).$$

Montrez que $R_n^\pi(\hat{\theta}) = \mathbb{E}\big[(\hat{\theta}(X_1, \ldots, X_n) - \theta)^2\big]$, où le vecteur aléatoire $(\theta, X_1, \ldots, X_n)$ est distribué selon la densité p.

2. **En déduire que le risque bayésien $R_n^\pi(\hat{\theta})$ est minimal en**

$$\hat{\theta}^\pi(x_1, \ldots, x_n) = \mathbb{E}\left\{\tilde{\theta} \,|\, \tilde{X}_1 = x_1, \ldots, \tilde{X}_n = x_n\right\} = \int \theta\, \pi(\theta | x_1, \ldots, x_n) d\theta$$

$$\textbf{où} \quad \pi(\theta | x_1, \ldots, x_n) = \frac{p(\theta, x_1, \ldots, x_n)}{\int p(\alpha, x_1, \ldots, x_n) d\alpha} \quad \textbf{(avec la convention } 0/0 = 0),$$

est la densité de la loi dite *loi a posteriori* de θ. L'estimateur $\hat{\theta}^\pi(X_1, \ldots, X_n)$ est l'estimateur bayésien de θ.

3. **On suppose maintenant que $f(x, \theta)$ est la densité de la loi $\mathcal{N}(\theta, \sigma^2)$ et la loi a priori est $\mathcal{N}(0, \sigma_0^2)$, où les variances σ^2 et σ_0^2 sont connues. Expliciter l'estimateur bayésien de θ. Quelle est sa forme limite dans les cas $n \to \infty$, $\sigma_0^2 \to 0$ et $\sigma_0^2 \to \infty$?**

****

7 Examen du lundi 26 octobre 2015

****

Exercice 7.1 (Estimation de la variance et borne de Cramer-Rao)

On considère le modèle d'échantillonnage $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathcal{N}(0, \theta)$ où $\theta > 0$ (la variance) est le paramètre inconnu à estimer.

1. **Calculer l'information de Fisher en $\theta > 0$ contenue dans ce n-échantillon.**
2. **Déterminer l'estimateur du maximum de vraisemblance $\widehat{\theta}_{\mathrm{n}}^{\mathtt{mv}}$ de θ.**
3. **Calculer le biais $b(\theta) = \mathbb{E}_\theta \widehat{\theta}_{\mathrm{n}}^{\mathtt{mv}} - \theta$ et le risque quadratique $R_\theta(\widehat{\theta}_{\mathrm{n}}^{\mathtt{mv}}) = \mathbb{E}_\theta(\widehat{\theta}_{\mathrm{n}}^{\mathtt{mv}} - \theta)^2$ de $\widehat{\theta}_{\mathrm{n}}^{\mathtt{mv}}$.**
4. **Rappeler la borne de Cramer-Rao pour ce problème. En déduire, que $\widehat{\theta}_{\mathrm{n}}^{\mathtt{mv}}$ atteint la borne de Cramer-Rao parmi tous les estimateurs sans biais.**

Rappel : si $g \sim \mathcal{N}(0, 1)$ alors $\mathbb{E}g^4 = 3$.

****
****

Exercice 7.2 (Estimateur on-line de la moyenne)

Dans le modèle d'échantillonnage $X_1, \ldots, X_n$ où $\mathbb{E}|X_1| < \infty$, on note $\mathbb{E}X_1 = \theta$; construire :

1. **un estimatuer *batch* de la moyenne θ**
2. **un estimateur *on-line* de la moyenne θ**

****
****

Exercice 7.3 (Deux échantillons gaussiens)

On observe $X_1, \ldots, X_m \overset{i.i.d.}{\sim} \mathcal{N}(\mu_1, v)$ **et** $Y_1, \ldots, Y_n \overset{i.i.d.}{\sim} \mathcal{N}(\mu_2, v)$ **deux échantillons Gaussiens ayant même variance** v **mais des moyennes différentes. On suppose que les deux échantillons sont indépendants entre eux.**

1. **Calculer la vraisemblance en** (μ_1, μ_2, v) **de l'observation** $(X_1, \ldots, X_m, Y_1, \ldots, Y_n)$**.**
2. **En déduire l'estimateur du maximum de vraisemblance de** (μ_1, μ_2, v)**.**
3. **On suppose dorénavant dans toutes les questions qui suivent que** $m = n$**. Calculer l'information de Fisher en** (μ_1, μ_2, v) **contenue dans le** n**-échantillon** $(X_1, Y_1), \ldots, (X_n, Y_n)$**.**
4. **On suppose que le modèle est régulier ; donner le comportement asymptotique de l'estimateur du maximum de vraisemblance.**
5. **Donner un test de niveau** α **consistant pour le problème de test**

$$H_0 : \mu_1 = 0 \text{ contre } H_1 : \mu_1 \neq 0$$

****
****

Exercice 7.4 (Ceinture de sécurité)

Une enquête sur l'influence de la ceinture de sécurité a donné les résultats suivants : sur 10.779 **conducteurs ayant subit un accident l'enquête rapporte les effectifs dans le tableau qui suit selon la gravité et le port au non de la ceinture de sécurité :**

nature des blessures	port de la ceinture	pas de ceinture
graves ou fatales	5	141
blessures sérieuses	25	330
peu ou pas de blessures	1229	9049

On souhaite répondre à la question : ***la ceinture de sécurité a-t'elle une influence sur la gravité des blessures lors d'un accident ?***

1. **Modéliser ces données.**
2. **Définir un problème de test permettant de répondre à la question.**
3. **Construire un test de niveau asymptotique** $\alpha = 0.05$**, consistant pour ce problème.**
4. **Comparer la p-value de ce test à** $0,001$**. Répondre à la question d'origine et donner un niveau de confiance sur votre décision.**

On rappel les quantiles d'ordre $1 - \alpha$ **d'une** $\chi^2(2)$ **:**

α	**0,999**	**0,995**	**0,99**	**0,98**	**0,95**	**0,9**	**0,8**	**0,2**	**0,1**
$q_{1-\alpha}^{\chi^2(2)}$	**0,0020**	**0,0100**	**0,0201**	**0,0404**	**0,1026**	**0,2107**	**0,4463**	**3,2189**	**4,6052**
α	**0,05**	**0,02**	**0,01**	**0,005**	**0,001**				
$q_{1-\alpha}^{\chi^2(2)}$	**5,9915**	**7,8240**	**9,2103**	**10,5966**	**13,8155**				

****

8 Rattrapage 2015-2016

****

Exercice 8.1 (Modèle d'uniforme perturbées)

Soit le modèle d'échantillonnage $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathbb{P}_\theta$ pour $\theta \in]-1,1[$ où $\mathbb{P}_\theta$ est une loi admettant une densité par rapport à la mesure de Lebesgue donnée par

$$f(\theta, x) = \frac{d\mathbb{P}_\theta}{d\lambda}(x) = (1-\theta)I(-1/2 < x < 0) + (1+\theta)I(0 < x < 1/2).$$

On pose

$$Y_n = \text{card}\{i : X_i > 0\} = \sum_{i=1}^{n} I(X_i > 0).$$

a) <u>Préliminaires</u>

1. **Donner l'expérience statistique associée à ces données.**
2. **Calculer $\mathbb{P}_\theta([0,1/2])$, la moyenne $\mathbb{E}_\theta X_1$ et la variance $\text{Var}(X_1)$.**
3. **Donner la loi de Y_n, sa moyenne et sa variance.**
4. **Vérifier que**
$$f(\theta, x) = (1-\theta)^{1-I(0<x<1/2)}(1+\theta)^{I(0<x<1/2)}.$$
En déduire l'expression de la vraisemblance de l'échantillon en θ en fonction de Y_n.
5. **Calculer l'information de Fisher sur θ contenue dans un n-échantillon de ce modèle.**

b) <u>Estimation de θ</u>

1. **Proposer un estimateur des moments de θ en fonction de Y_n.**
2. **Montrer que l'estimateur du maximum de vraisemblance vaut $\widehat{\theta}_{\text{n}}^{\text{mv}} = \frac{2}{n}Y_n - 1$.**
3. **Etudier les propriétés de $\widehat{\theta}_{\text{n}}^{\text{mv}}$: biais, variance, consistance.**
4. **Comparer le risque quadratique de $\widehat{\theta}_{\text{n}}^{\text{mv}}$ et la borne de Cramer-Rao. En déduire que $\widehat{\theta}_{\text{n}}^{\text{mv}}$ atteint la borne de Cramer-Rao parmi tous les estimateurs sans biais.**
5. **Montrer que sous $\mathbb{P}_\theta$, $\sqrt{n}\big(\widehat{\theta}_{\text{n}}^{\text{mv}} - \theta\big)$ converge en loi vers $\mathcal{N}(0, 1-\theta^2)$.**
6. **Etudier le comportement asymptotique de**
$$\frac{\sqrt{n}\big(\widehat{\theta}_{\text{n}}^{\text{mv}} - \theta\big)}{\sqrt{1 - \widehat{\theta}_{\text{n}}^{\text{mv}^2}}}.$$
7. **Construire un intervalle de confiance pour θ de niveau asymptotique $\alpha = 0.95$ centré en $\widehat{\theta}_{\text{n}}^{\text{mv}}$ et de longueur proportionnelle à $n^{-1/2}$.**

c) <u>Tests</u>

1. **On considère le problème de test :**

$$H_0 : \theta = 0 \text{ contre } H_1 : \theta = 1/2$$

Sous quelles condition existe-t'il un test de Neyman-Pearson de niveau α (on ne considère ici que les tests non randomizés). Dans ce cas, existe-t'il un test de même niveau plus puissant ?

2. **Pour le même problème de test, construire un test de niveau asymptotique α. Etudier sa puissance.**
3. **On considère le problème de test :**

$$H_0 : \theta = 0 \text{ contre } H_1 : \theta \neq 0$$

Construire un test de niveau asymptotique α. Etudier sa consistance.

d) Application

On considère un n-échantillon $U_1, \ldots, U_n \overset{i.i.d.}{\sim} \mathcal{U}([-1/2, 1/2])$. Un phénomène aléatoire perturbe les observations des U_i : pour chaque $i = 1, \ldots, n$, la quantité $|U_i|$ est observée avec probabilité $\theta \in [0, 1)$ sinon c'est U_i qui est observée. Ces perturbations sont indépendantes entres elles et indépendantes des U_i. On note $X_1, \ldots, X_n$ l'échantillon finalement observé après perturbation.

1. **Déterminer la loi de X_1.**
2. **Proposer une méthode d'estimation de θ.**
3. **Construire un test de niveau asymptotique α consistant permettant de décider si un tel phénomène de perturbation s'est produit.**
4. **La loi des U_i n'étant plus uniforme, que suffit-il de connaître sur elle pour que ce test reste valable ?**

****