AP 19

Estimation

Exercice 1 (ECRICOME)

Soient a et b deux réels tels que $a \geq 0$ et $b > 0$. On considère la fonction $f_{a,b}$ définie sur $\mathbb{R}$ par :

$$f_{a,b}(x) = \begin{cases} \dfrac{1}{b}\exp\left(-\dfrac{x-a}{b}\right) & \text{si } x \geq a, \\ 0 & \text{sinon.} \end{cases}$$

1. Vérifier que $f_{a,b}$ est bien une densité de variable aléatoire. On note $\mathcal{E}(a,b)$ la loi associée.
2. On considère désormais une variable aléatoire X de loi $\mathcal{E}(a,b)$.

   Déterminer la fonction de répartition de X.
3. On pose $Y = X - a$. Déterminer la loi de Y et la reconnaitre. En déduire $E(X)$ et $V(X)$.

On considère désormais une suite de variables aléatoires $(X_i)_{i\geq 1}$ indépendantes, identiquement distribuées de loi $\mathcal{E}(a,b)$, c'est-à-dire qui suivent toutes la loi $\mathcal{E}(a,b)$.
Pour tout entier n supérieur ou égal à 2, on considère les variables aléatoires S_n et Y_n définies par $S_n = X_1 + \ldots + X_n$ et $Y_n = \min(X_1, \ldots, X_n)$.
Le but des questions suivantes est de déterminer des estimateurs de a et b.

4. Déterminer l'espérance et la variance de S_n.
5. (a) Déterminer la fonction de répartition de Y_n.
   (b) En déduire que Y_n suit une loi $\mathcal{E}(a_n, b_n)$ (on précisera a_n et b_n).
   (c) Donner les valeurs de $E(Y_n)$ et $V(Y_n)$.
6. On pose $Z_n = \dfrac{S_n}{n} - Y_n$.
   (a) Montrer que $\lim\limits_{n\to+\infty} E(Z_n) = b$.
   (b) On note $r_b(Z_n) = E((Z_n - b)^2)$ le risque quadratique de Z_n. Montrer que :
   $$r_b(Z_n) = \frac{b^2}{n^2} + V(Z_n).$$
   (c) En déduire que :
   $$r_b(Z_n) = \frac{2b^2}{n^2} + \frac{b^2}{n} - \frac{2}{n}Cov(S_n, Y_n).$$
   (d) Justifier que $(Cov(S_n, Y_n))^2 \leq V(S_n)V(Y_n)$, puis montrer que $\lim\limits_{n\to+\infty} r_b(Z_n) = 0$.

Exercice 2 (EML)

Dans cet exercice, toutes les variables aléatoires sont supposées définies sur un même espace probabilisé noté $(\Omega, \mathcal{A}, P)$.

Partie A : Loi de Pareto

Soient a et b deux réels strictement positifs. On définit la fonction f sur $\mathbb{R}$ par :

$$f : x \mapsto \begin{cases} 0 & \text{si } x < b, \\ a\dfrac{b^a}{x^{a+1}} & \text{si } x \geq b. \end{cases}$$

1. Montrer que f est une densité de probabilité.

On dit qu'une variable aléatoire suit la loi de Pareto de paramètres a et b lorsqu'elle admet pour densité la fonction f.
Dans toute la suite de l'exercice, on considère une variable aléatoire X suivant la loi de Pareto de paramètres a et b.

2. Déterminer la fonction de répartition de X.

3. (a) Soit U une variable aléatoire suivant la loi uniforme sur $]0,1]$.
   Montrer que la variable aléatoire $b\,U^{-\frac{1}{a}}$ suit la loi de Pareto de paramètres a et b.

   (b) En déduire une fonction `Python` d'en-tête `def pareto(a,b)` qui prend en arguments deux réels a et b strictement positifs et qui renvoie une simulation de la variable aléatoire X.

   (c) On considère la fonction `Python` ci-dessous.
   Que contient la liste `L` renvoyée par la fonction `mystere` ?

```python
def mystere(a, b)
    L = []
    for p in range(2, 7):
        S = 0
        for k in range(10**p):
            S = S + pareto(a,b)
        L.append(S/10**p)
    return(L)
```

   (d) On exécute la fonction précédente avec différentes valeurs de a et de b.
   Comment interpréter les résultats obtenus ?

```
>>>mystere(2,1)
  [1.9306917  1.9411352  1.9840089  1.9977684  2.0012415]
>>>mystere(3,2)
  [3.1050951  3.0142956  2.9849407  2.9931656  2.9991517]
>>>mystere(1,4)
  [21.053151  249.58609  51.230522  137.64549  40.243918]
```

4. (a) Montrer que X admet une espérance si et seulement si $a > 1$ et que, dans ce cas :

$$E(X) \;=\; \frac{a\,b}{a-1}$$

   (b) Montrer que X admet une variance si et seulement si $a > 2$ et que, dans ce cas :

$$V(X) \;=\; \frac{a\,b^2}{(a-1)^2\,(a-2)}$$

Partie B : Estimation du paramètre b
On suppose **dans cette partie uniquement** que $a = 3$ et on chercher à déterminer un estimateur performant de b. Ainsi, la variable aléatoire X admet pour densité la fonction f définie par :

$$f : x \mapsto \begin{cases} 0 & \text{si } x < b, \\ \dfrac{3b^3}{x^4} & \text{si } x \geq b. \end{cases}$$

Soient $n \in \mathbb{N}^*$ et $X_1, \ldots, X_n$ des variables aléatoires indépendantes, toutes de même loi que X.
On définit :

$$Y_n = \min(X_1, \ldots, X_n) \qquad \text{et} \qquad Z_n = \frac{X_1 + \cdots + X_n}{n}$$

On admet que Y_n et Z_n sont encore des variables aléatoires définies sur $(\Omega, \mathcal{A}, P)$.

5. (a) Calculer, pour tout x de $[b, +\infty[$, $P(Y_n > x)$.
   (b) En déduire que Y_n suit une loi de Pareto dont on précisera les paramètres.
   (c) Montrer que $Y'_n = \dfrac{3n-1}{3n} Y_n$ admet une espérance égale à b.
   Calculer la variance de Y'_n.

6. (a) Déterminer l'espérance et la variance de Z_n.
   (b) En déduire un estimateur noté Z'_n d'espérance égale à b de la forme $\alpha\, Z_n$ où α est un réel à préciser.
   Calculer la variance de Z'_n.

7. Entre Y'_n et Z'_n, quel estimateur choisir ? Justifier.

Partie C : Estimation du paramètre a
On suppose **dans cette partie uniquement** que $b = 1$ et on cherche à construire un intervalle de confiance pour a. Ainsi, la variable aléatoire X admet pour densité la fonction f définie par :

$$f : x \mapsto \begin{cases} 0 & \text{si } x < 1, \\ \dfrac{a}{x^{a+1}} & \text{si } x \geq 1. \end{cases}$$

Soit $(X_n)_{n\in\mathbb{N}^*}$ une suite de variables aléatoires indépendantes, toutes de même loi que X.

8. Soit $n \in \mathbb{N}^*$. On pose : $W_n = \ln(X_n)$.
   Montrer que la variable aléatoire W_n suit une loi exponentielle dont on précisera le paramètre.
   En déduire l'espérance et la variance de W_n.

9. On définit, pour tout n de $\mathbb{N}^*$:

$$M_n = \frac{\ln(X_1) + \cdots + \ln(X_n)}{n} \qquad \text{et} \qquad T_n = \sqrt{n}\,(a\, M_n - 1)$$

   (a) Justifier que la suite de variables aléatoires $(T_n)_{n\in\mathbb{N}^*}$ converge en loi vers une variable aléatoire suivant la loi normale centrée réduite.
   (b) En déduire que l'intervalle $\left[\dfrac{\sqrt{n}-2}{\sqrt{n}M_n} \; ; \; \dfrac{\sqrt{n}+2}{\sqrt{n}M_n}\right]$ est un intervalle de confiance asymptotique pour a au niveau de confiance 95%.
   On admettra que $\Phi(2) \geq 0,975$, où Φ désigne la fonction de répartition d'une variable aléatoire suivant la loi normale centrée réduite.

Exercice 3 (ESSEC)

Notations

- Tout au long du sujet $(\Omega, \mathcal{F}, P)$ désignera un espace probabilisé et les variables aléatoires utilisées seront toutes définies sur cet espace probabilisé. Sous réserve d'existence, l'espérance mathématique d'une variable aléatoire réelle X sera notée $E(X)$ et sa variance sera notée $V(X)$.

- Pour un événement A, on notera $P_B(A)$ la probabilité conditionnelle de A sachant B où B est un événement non négligeable.

Le sujet est composé de quatre parties. Les parties I, II, III et IV.A sont **indépendantes**.
Il s'agit de variations autour de la notion de risque quadratique en théorie de l'estimation.

I. Premier problème d'estimation

Dans ce premier problème d'estimation, on dispose d'une seule observation notée X. On suppose que X admet pour densité f_θ définie sur $\mathbb{R}$ par

$$f_\theta(x) = \begin{cases} \dfrac{k+1}{\theta^{k+1}}\, x^k & \text{pour tout } x \in [0,\theta] \\ 0 & \text{sinon} \end{cases}$$

où k est un entier naturel non nul et θ un paramètre réel inconnu strictement positif que l'on souhaite estimer.

1. Montrer que f_θ est bien une densité de probabilité.
2. Calculer $E(X)$.

On définit le biais de T estimateur de θ par :

$$b(T,\theta) = E(T-\theta)$$

On dira que T est sans biais si $b(T,\theta) = 0$ et qu'il est biaisé sinon.

3. Déterminer λ_0 réel dépendant uniquement de k tel que $\lambda_0 X$ estime θ sans biais.
4. Calculer $V(X)$.

On définit le risque quadratique de T estimateur de θ par :

$$R(T,\theta) = E\left((T-\theta)^2\right)$$

5. Démontrer que pour tout T estimateur de θ, on a :

$$R(T,\theta) = (b(T,\theta))^2 + V(T)$$

6. Donner la valeur de $R(\lambda_0 X, \theta)$.

Le but de la fin de cette partie I est de déterminer un estimateur de θ ayant un plus petit risque quadratique que celui de $\lambda_0 X$.

7. En utilisant I.5 montrer que pour tout λ réel

$$R(\lambda X, \theta) = \theta^2 Q(\lambda)$$

   où Q est un polynôme de degré 2 dont les coefficients ne dépendent que de k.
8. Montrer que la fonction $\lambda \mapsto Q(\lambda)$ atteint son minimum en un unique réel noté λ^* que l'on exprimera en fonction de k.
9. Conclure sur le but recherché.

II. Second problème d'estimation

Dans ce second problème d'estimation, on dispose de n observations indépendantes ($n \geq 2$) notées $X_1, ..., X_n$ de même loi de Poisson de paramètre θ inconnu ($\theta \in]0, +\infty[$). On souhaite estimer le paramètre $\exp(-\theta)$.
On définit pour tout i élément de $\{1, ..., n\}$ la variable aléatoire Y_i par

$$Y_i = 1 \text{ si } X_i = 0 \quad \text{et} \quad Y_i = 0 \text{ sinon.}$$

Puis on note

$$\overline{Y_n} = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

10. Pour tout i élément de $\{1, ..., n\}$, donner la loi de Y_i.

11. Donner la loi de $\sum_{i=1}^{n} Y_i$, puis montrer que $E\left(\overline{Y_n}\right) = \exp(-\theta)$.

    On dira dans ce cas que $\overline{Y_n}$ est un estimateur sans biais de $\exp(-\theta)$.

12. Calculer $V\left(\overline{Y_n}\right)$.

Pour tout k élément de $\{1, ..., n\}$ on définit $S_k = \sum_{i=1}^{k} X_i$.

13. Rappeler sans démonstration la loi de S_k pour tout k élément de $\{1, ..., n\}$.

On définit jusqu'à la fin de cette partie II pour tout j entier naturel

$$\varphi(j) = P_{\{S_n = j\}}(X_1 = 0)$$

14. Montrer que pour tout j entier naturel

$$\varphi(j) = \left(1 - \frac{1}{n}\right)^j$$

    On a donc $\varphi(j)$ indépendant du paramètre θ inconnu.

D'après la question 14, on peut définir l'estimateur

$$\varphi(S_n) = \left(1 - \frac{1}{n}\right)^{S_n}$$

15. Montrer que $\varphi(S_n)$ admet une espérance et que $E(\varphi(S_n)) = \exp(-\theta)$.

    On dira dans ce cas que $\varphi(S_n)$ est un estimateur sans biais de $\exp(-\theta)$.

16. Montrer que $\varphi(S_n)$ admet une variance vérifiant

$$V(\varphi(S_n)) = \exp(-2\theta)\left(\exp\left(\frac{\theta}{n}\right) - 1\right)$$

17. On souhaite comparer les performances de $\overline{Y_n}$ et $\varphi(S_n)$ en tant qu'estimateurs de $\exp(-\theta)$.

    (a) En utilisant le théorème des accroissements finis, démontrer que

$$1 \leq \frac{\exp(\theta) - 1}{\theta} \leq \exp(\theta)$$

    (b) Soit la fonction $h : [0, 1] \to \mathbb{R}$ définie par

$$h(t) = t\exp(\theta) + (1 - t) - \exp(t\theta)$$

    pour tout $t \in [0, 1]$. Étudier les variations de h.

    (c) En déduire que

$$\exp\left(\frac{\theta}{n}\right) \leq \frac{\exp(\theta)}{n} + \frac{n-1}{n}$$

    puis l'inégalité

$$V(\varphi(S_n)) \leq V\left(\overline{Y_n}\right)$$

(d) On définit le risque quadratique de T_n estimateur de $\exp(-\theta)$ par

$$R(T_n, \exp(-\theta)) = E\left((T_n - \exp(-\theta))^2\right)$$

Comparer les risques quadratiques de $\overline{Y_n}$ et $\varphi(S_n)$.

On reprendra à la fin de la partie IV l'étude de $\varphi(S_n)$.

III. Information de Fisher

A. Cas discret

Dans cette section III.A, on considère I un intervalle de $\mathbb{R}$, θ un paramètre inconnu de I et X une variable aléatoire à valeurs dans $\mathbb{N}$ ($X(\Omega) \subseteq \mathbb{N}$). On suppose qu'il existe une fonction p définie sur $I \times X(\Omega)$ telle que pour tout k élément de $X(\Omega)$

$$P(X = k) = p(\theta, k)$$

et vérifiant pour tout k de $X(\Omega)$, $\theta \mapsto p(\theta, k)$ dérivable sur I.
On définit sous réserve d'existence l'**information de Fisher** de X par

$$I_X(\theta) = \sum_{k \in X(\Omega)} \left(\frac{\partial}{\partial \theta} \ln p(\theta, k)\right)^2 p(\theta, k)$$

18. Dans cette question, on considère X variable aléatoire de loi de Bernoulli de paramètre θ ($\theta \in]0, 1[$).

    On a alors $X(\Omega) = \{0, 1\}$, $P(X = 1) = p(\theta, 1) = \theta$, $P(X = 0) = p(\theta, 0) = 1 - \theta$ et

    $$I_X(\theta) = \left(\frac{\partial}{\partial \theta} \ln p(\theta, 1)\right)^2 p(\theta, 1) + \left(\frac{\partial}{\partial \theta} \ln p(\theta, 0)\right)^2 p(\theta, 0)$$

    Montrer que

    $$I_X(\theta) = \frac{1}{\theta(1 - \theta)}$$

19. Dans cette question, on considère X variable aléatoire de loi binomiale de paramètres N et θ ($N \in \mathbb{N}^*$, $\theta \in]0, 1[$).

    (a) Montrer que

    $$I_X(\theta) = \frac{1}{(\theta(1 - \theta))^2} \sum_{k=0}^{N} (k - N\theta)^2 \binom{N}{k} \theta^k (1 - \theta)^{N-k}$$

    (b) En déduire que

    $$I_X(\theta) = \frac{V(X)}{(\theta(1 - \theta))^2}$$

    puis donner la valeur de $I_X(\theta)$.

20. Dans cette question, on considère X une variable aléatoire de Poisson de paramètre θ ($\theta \in]0, +\infty[$). Puisque $X(\Omega) = \mathbb{N}$, on a sous réserve de convergence

    $$I_X(\theta) = \sum_{k=0}^{+\infty} \left(\frac{\partial}{\partial \theta} \ln p(\theta, k)\right)^2 p(\theta, k)$$

    (a) Montrer que la série de terme général $\left(\frac{\partial}{\partial \theta} \ln p(\theta, k)\right)^2 p(\theta, k)$ converge et calculer sa somme $I_X(\theta)$.

(b) Justifier que

$$I_X(\theta) = E\left(\left(\frac{\partial}{\partial \theta}\ln p(\theta, X)\right)^2\right).$$

B. Cas d'une variable gaussienne

Soit X une variable aléatoire qui suit la loi normale de moyenne θ ($\theta \in \mathbb{R}$) et de variance 1 dont la densité est notée $x \mapsto f(\theta, x)$. On définit sous réserve de convergence l'**information de Fisher** de X par

$$I_X(\theta) = \int_{-\infty}^{+\infty}\left(\frac{\partial}{\partial \theta}\ln(f(\theta, x))\right)^2 f(\theta, x)\,dx$$

21. Montrer que sous réserve de convergence

$$I_X(\theta) = \int_{-\infty}^{+\infty}(x-\theta)^2 f(\theta, x)\,dx$$

22. En déduire l'existence et la valeur de $I_X(\theta)$.

23. Justifier que

$$I_X(\theta) = E\left(\left(\frac{\partial}{\partial \theta}\ln(f(\theta, X))\right)^2\right)$$

IV. Minoration du risque quadratique

A. Inégalité de Cramer-Rao

Dans cette section IV.A, on considère I un intervalle de $\mathbb{R}$, θ un paramètre inconnu de I et X une variable aléatoire telle que $X(\Omega) = \{0, ..., N\}$ ($N \in \mathbb{N}$). On suppose qu'il existe une fonction p définie sur $I \times X(\Omega)$ telle que pour tout $k \in \{0, ..., N\}$

$$P(X = k) = p(\theta, k)$$

et vérifiant :

- pour tout $k \in \{0, ..., N\}$, $\theta \mapsto p(\theta, k)$ dérivable sur I,
- l'information de Fisher de X notée $I_X(\theta)$ définie dans la partie III est non nulle pour tout $\theta \in I$.

Le but de la section IV.A est de démontrer l'inégalité suivante due à Cramer et Rao.

Théorème de Cramer-Rao
Soit $f(X)$ un estimateur sans biais de $g(\theta)$ à savoir tel que $E(f(X)) = g(\theta)$ où g est dérivable sur I. On a alors

$$V(f(X)) \geq \frac{(g'(\theta))^2}{I_X(\theta)}$$

24. Montrer que pour tout θ élément de I,

$$\sum_{k=0}^{N}\frac{\partial}{\partial \theta}(p(\theta, k)) = 0$$

25. En déduire que pour tout θ élément de I

$$E\left(\frac{\partial}{\partial \theta}\ln(p(\theta, X))\right) = 0 \qquad \text{(E)}$$

26. En dérivant partiellement par rapport à θ les deux membres de l'égalité (E), montrer que pour tout θ élément de I,

$$E\left(\frac{\partial^2}{\partial\theta^2}\ln\left(p\left(\theta,X\right)\right)\right) = -E\left(\left(\frac{\partial}{\partial\theta}\ln\left(p\left(\theta,X\right)\right)\right)^2\right).$$

27. Montrer que pour tout θ élément de I,

$$g'\left(\theta\right) = \sum_{k=0}^{N} f\left(k\right)\left(\frac{\partial}{\partial\theta}\ln\left(p\left(\theta,k\right)\right)\right)\ p\left(\theta,k\right)$$

puis que

$$g'\left(\theta\right) = E\left(\left(f\left(X\right)-g\left(\theta\right)\right)\frac{\partial}{\partial\theta}\ln\left(p\left(\theta,X\right)\right)\right)$$

28. On pose pour tout t réel

$$L\left(t\right) = E\left(\left(\left(f\left(X\right)-g\left(\theta\right)\right)+t\frac{\partial}{\partial\theta}\ln\left(p\left(\theta,X\right)\right)\right)^2\right)$$

   (a) Développer le polynôme L suivant les puissances décroissantes de t.
   (b) Calculer le discriminant Δ de L et justifier que $\Delta \leq 0$.
   (c) En déduire l'inégalité de Cramer-Rao.

B. Extension du théorème de Cramer-Rao

On reprend dans cette section IV.B les notations et hypothèses de la partie II. On admet que, dans ce contexte, le théorème de Cramer-Rao peut se généraliser comme suit :

Théorème de Cramer-Rao
Soit $T_n = f\left(X_1,...,X_n\right)$ un estimateur sans biais de $g\left(\theta\right)$ à savoir tel que $E\left(f\left(X_1,...,X_n\right)\right) = g\left(\theta\right)$ où g est dérivable sur $]0,+\infty[$. On a alors

$$V\left(T_n\right) \geq \frac{\left(g'\left(\theta\right)\right)^2}{nI_{X_1}\left(\theta\right)}$$

où $I_{X_1}\left(\theta\right)$ est l'information de Fisher d'une variable aléatoire de loi de Poisson de paramètre θ définie et calculée à la partie III.

Il s'agit dans cette section d'exploiter cette nouvelle inégalité de Cramer-Rao. On note

$$\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

29. Calculer $E\left(\overline{X_n}\right)$ et $V\left(\overline{X_n}\right)$.

30. Déduire de la généralisation de Cramer-Rao, que $\overline{X_n}$ a le plus petit risque quadratique parmi les estimateurs sans biais de θ.

31. Montrer que pour $g\left(\theta\right) = \exp\left(-\theta\right)$ où $\theta \in]0,+\infty[$

$$V\left(\varphi\left(S_n\right)\right) \underset{n\to+\infty}{\sim} \frac{\left(g'\left(\theta\right)\right)^2}{nI_{X_1}\left(\theta\right)}$$

32. Que prouve ce résultat en terme d'optimalité de $\varphi\left(S_n\right)$ dans l'estimation de $\exp\left(-\theta\right)$?

33. A la lumière de la partie I, peut-on conclure que lorsque n est grand $\varphi\left(S_n\right)$ est le meilleur estimateur de $\exp\left(-\theta\right)$ en terme de risque quadratique ?