TD19

Estimation ponctuelle

Exercice 19.1 (★)

Soient $X_1, \dots, X_n$ des variables aléatoires i.i.d. telles que :

$$P(X_1 = -1) = (1-p)^2,\ P(X_1 = 0) = 2p(1-p),\ P(X_1 = 1) = p^2.$$

On cherche à estimer $p \in]0,1[$. Montrer que $Z_n = \displaystyle\sum_{k=1}^{n} \frac{1+X_k}{2n}$ est un estimateur sans biais convergent de p.

Exercice 19.2 (★★)

Soit T une variable aléatoire qui suit la loi normale centrée et de variance σ^2 inconnue, $\sigma > 0$. Pour tout entier $n \geq 2$, on dispose d'un n-échantillon i.i.d. $(T_1, T_2, \dots, T_n)$ de la loi de T.

On considère la variable aléatoire S_n définie par $S_n = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} T_i^2$.

Montrer que S_n est un estimateur sans biais et convergent du paramètre σ^2.

Exercice 19.3 (★★)

Soit $(X_n)_{n\geq 1}$ une suite de variables aléatoires indépendantes suivant la loi exponentielle de paramètre $\lambda > 0$. On pose $S_n = \displaystyle\sum_{i=1}^{n} X_i$.

1. Déterminer une densité de S_n.
2. Calculer l'espérance de $\dfrac{1}{S_n}$. En déduire un estimateur sans biais de λ.

Exercice 19.4 (★★ - Variance empirique - ✍)

On suppose que la loi mère d'un échantillon $(X_1, \dots, X_n)$ possède une espérance m et une variance $\sigma^2 > 0$.

1. On suppose dans cette question que m connu, et on pose :

$$T_n = \sqrt{\frac{1}{n}\sum_{k=1}^{n}(X_k - m)^2}.$$

   Montrer que T_n^2 est un estimateur sans biais de σ^2.

2. On suppose maintenant que m est inconnu et on pose :

$$S_n = \sqrt{\frac{1}{n}\sum_{k=1}^{n}(X_k - \overline{X_n})^2}.$$

   (a) Montrer que $S_n^2 = \dfrac{1}{n}\left(\displaystyle\sum_{k=1}^{n} X_k^2\right) - \overline{X_n}^2$.

   (b) Montrer que S_n^2 est un estimateur asymptotiquement sans biais de σ^2.
   (c) Donner un estimateur sans biais de σ^2.

3. On suppose enfin que la loi mère admet un moment d'ordre 4. On admet que si une suite (X_n) converge en probabilité vers X et (Y_n) converge en probabilité vers Y, alors $X_n + Y_n \xrightarrow{P} X + Y$.
   (a) En utilisant les théorèmes opératoires sur la convergence en probabilité, montrer que T_n^2 et S_n^2 sont des estimateurs convergents de σ^2.
   (b) Conclure que T_n et S_n sont des estimateurs convergents de σ.

Exercice 19.5 (★★)

On dispose d'un n-échantillon $(X_1, X_2, \ldots, X_n)$ de variables aléatoires indépendantes suivant la loi de Poisson de paramètre λ avec λ inconnu.
On cherche à estimer la probabilité $e^{-q\lambda}$ de n'observer que des 0 au cours de q expériences consécutives.
On pose $T_n = e^{-q\overline{X_n}}$.
T_n est-il un estimateur sans biais de $e^{-q\lambda}$? Asymptotiquement sans biais ? Convergent ?

Exercice 19.6 (★★)

On effectue un sondage pour estimer le résultat d'une élection. Notons pour cela q la proportion d'individus déclarant vouloir voter pour le candidat A, $1-q$ la proportion d'individus déclarant vouloir voter pour le candidat B.

1. Quel estimateur de q souhaiteriez vous utiliser ? Justifier votre choix.
2. On estime que 5 personnes interrogées sur 6 répondent honnêtement au sondeur, une personne sur 6 disant le contraire de ce qu'elle pense vraiment. On note p la proportion d'individus votant effectivement pour A, et $1-p$ celle votant pour B.
   (a) Exprimer la probabilité q qu'une personne réponde « Candidat A » au sondeur en fonction de p.
   (b) Déterminer un estimateur convergent de p.

Exercice 19.7 (★)

Soit $(X_n)_{n\geq 1}$ une suite de variables aléatoires indépendantes suivant la loi exponentielle $\mathcal{E}(\lambda)$. On pose :

$$\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{et} \quad T_n = \frac{1}{n+1}\sum_{i=1}^{n} X_i.$$

Comparer les risques quadratiques de $\overline{X_n}$ et T_n en tant qu'estimateurs de $\dfrac{1}{\lambda}$.

Exercice 19.8 (★★)

On cherche à estimer le paramètre p d'une loi de Bernoulli. Soit $(X_1, \ldots, X_n)$ un n-échantillon de loi mère $\mathscr{B}(p)$, et notons $S_n = \displaystyle\sum_{i=1}^{n} X_i$. On pose :

$$\overline{X_n} = \frac{S_n}{n} \quad \text{et} \quad T_n = \frac{S_n + 1}{n+2}.$$

Comparer les risques quadratiques de $\overline{X_n}$ et T_n en tant qu'estimateurs de p (on étudiera les cas où p est proche de 1 et de $\frac{1}{2}$). Peut-on privilégier l'un de ces estimateurs par rapport à l'autre ?

Exercice 19.9 (★★)

Soit θ un réel strictement positif, et soit f_θ la fonction définie sur $\mathbb{R}$ par :

$$f_\theta(x) = \begin{cases} 0 & \text{si } x \leq \theta \\ e^{-(t-\theta)} & \text{si } x > \theta \end{cases}$$

1. Montrer que f_θ est une densité de probabilité.
2. Soit X une variable aléatoire dont la densité est f_θ. Reconnaître la loi de $X - \theta$ et en déduire F_X, $E(X)$ et $V(X)$.
3. Dans toute la suite, $(X_n)_{n\geq 1}$ est une suite de variables aléatoires indépendantes sur $(\Omega, \mathcal{A})$, de densité f_θ. Pour tout $n \geq 1$, on pose

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

   (a) À partir de $\overline{X_n}$, construire un estimateur T_n sans biais et convergent de θ.

   (b) i. Déterminer la loi de Y_n.

   ii. Montrer que Y_n est un estimateur asymptotiquement sans biais de θ.

   iii. Construire à partir de Y_n un estimateur U_n sans biais et convergent de θ.

   (c) Comparer les estimateurs T_n et U_n.

Exercice 19.10 (★★ - Estimation du paramètre d'une loi de Poisson)

1. On considère que le nombre de poissons capturés lors d'une journée de pêche dans un lac suit une loi de Poisson de paramètre $\theta > 0$ que l'on cherche à évaluer.
   En nombre de prises, le palmarès de 15 journées est le suivant :

$$1, 6, 4, 2, 3, 5, 2, 7, 5, 3, 1, 4, 5, 3, 6.$$

   (a) Proposer un modèle statistique adapté à cette expérience.

   (b) Proposer deux estimateurs sans biais de θ.

   (c) À combien évaluez-vous le nombre moyen de poissons pêchés au cours d'une journée ?

2. On souhaite juger de la pertinence de la moyenne empirique et de la variance empirique (non biaisée) en tant qu'estimateurs du paramètre λ d'une loi de Poisson. Prenons pour cela `lbd = grand(1,1,'exp',1)` un nombre réel positif aléatoire, et $n = 15$ pour commencer.

   (a) Créer un n-échantillon de la loi $\mathscr{P}(\lambda)$ à l'aide de `Scilab`, et calculer les estimations obtenues de λ à l'aide de la moyenne empirique et de la variance empirique (non biaisée).

   (b) On répète cette procédure pour 1000 échantillons de taille $n = 15$. Écrire un programme qui :

   - simule 1000 n-échantillons i.i.d de la loi $\mathscr{P}(\lambda)$ (c'est-à-dire crée 1000 échantillons observés),

- calcule, pour chaque échantillon observé, l'estimation correspondante pour la moyenne empirique et pour la variance empirique (non biaisée),
- trace l'histogramme des 1000 estimations obtenues avec chacun des estimateurs.

(c) Comparer les deux histogrammes (on pourra utiliser les commandes `subplot(1,2,1)` et `subplot(1,2,2)` pour tracer les histogrammes sur la même fenêtre graphique l'un à côté de l'autre). Quel semble être le meilleur estimateur ?

(d) Que constate-t-on lorsque n devient grand ?

Exercice 19.11 (★★ - Estimateur du maximum de vraisemblance - ESSEC ECE 2007 - ✍)

1. Pour $a > 0$, montrer que la fonction :

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

est une densité de probabilité.

Dans la suite de cet exercice, on considère un n-échantillon $(X_1, \dots, X_n)$ i.i.d. de loi à densité f_a, où $a > 0$ est un paramètre inconnu.

2. On définit la *fonction de vraisemblance*

$$L : (x_1, \dots, x_n, a) \in \mathbb{R}^n \times \mathbb{R}_+^* \mapsto L(x_1, \dots, x_n, a) = \prod_{k=1}^{n} f_a(x_k).$$

(a) Expliciter, pour $(x_1, \dots, x_n, a) \in \mathbb{R}^n \times \mathbb{R}_+^*$, la valeur de $L(x_1, \dots, x_n, a)$.

(b) Pour $(x_1, \dots, x_n) \in [1, +\infty[^n$ donné, étudier les variations de la fonction $h : a \mapsto L(x_1, \dots, x_n, a)$ et montrer qu'elle atteint un maximum en un point $\hat{a}$. On pourra considérer la fonction $g : a \mapsto \ln(h(a))$.

Puisque $\hat{a}$ dépend de $x_1, \dots, x_n$, il existe une fonction $\varphi :]0, +\infty[^n \mapsto \mathbb{R}$ telle que $\hat{a} = \varphi(x_1, \dots, x_n)$. On considère alors la variable aléatoire $T_n = \varphi(X_1, \dots, X_n)$, appelée *estimateur du maximum de vraisemblance de a*.

3. (a) Pour $k \in [\![1, n]\!]$, montrer que la variable aléatoire $Y_k = \ln(X_k)$ suit une loi exponentielle.

(b) En déduire une densité de $S_n = Y_1 + \cdots + Y_n$.

(c) Calculer $E(T_n)$ et $V(T_n)$.

(d) En déduire que T_n est un estimateur asymptotiquement sans biais et convergent de a.

Exercice 19.12 (★★ - Estimation des paramètres d'une loi binomiale)

Soient $(X_k)_{k\in\mathbb{N}^*}$ une suite de variables aléatoires mutuellement indépendantes, suivant la loi binomiale $\mathcal{B}(N, p)$, où N et p sont inconnus. On cherche donc à estimer le paramètre $\theta = (N, p) \in \mathbb{N}^* \times]0, 1[$.

1. Pour tout $n \in \mathbb{N}^*$, on pose $M_n = \max(X_1, \dots, X_n)$.

(a) Calculer $P(M_n < N)$.

(b) Montrer que si $\varepsilon > 0$, alors $P(|M_n - N| > \varepsilon) \leq P(M_n < N)$.

(c) En déduire que M_n est un estimateur convergent de N.

2. On pose à présent $\overline{X_n} = \dfrac{X_1 + \cdots + X_n}{n}$. Montrer que $\overline{X_n}$ est un estimateur convergent de Np.

3. Dans cette question, on admet le résultat suivant :

   Si c est une constante, et (X_n) une suite de variables aléatoires, alors $X_n \xrightarrow{P} c$ ssi $X_n \xrightarrow{\mathscr{L}} c$.

   Montrer que $\dfrac{\overline{X_n}}{M_n}$ est un estimateur convergent de p.

1. Notons que M_n est une variable aléatoire discrète de support $[\![0, N]\!]$

   (a) On a :

   $$P(M_n < N) = P(M_n \leq N-1) = P\left(\bigcap_{k=1}^{n} [X_k \leq N-1]\right) \underset{X_k \text{ i.i.d.}}{=} P(X_1 \leq N-1)^n$$
   $$= (1 - P(X_1 = N))^n = (1-p^N)^n$$

   (b) Soit $\varepsilon > 0$. On a :

   $$P(|M_n - N| > \varepsilon) \underset{M_n \leq N}{=} P(N - M_n > \varepsilon) = P(N - \varepsilon > M_n) \leq P(M_n < N)$$

   car $[M_n < N] \subset [M_n < N - \varepsilon]$.

   (c) D'après les question précédentes, on a donc pour tout $\varepsilon > 0$:

   $$0 \leq P(|M_n - N| > \varepsilon) \leq P(M_n < N) = (1-p^N)^n.$$

   Puisque $p \in]0,1[$, on a $1 - p^N \in]0,1[$ et donc $\lim\limits_{n\to+\infty} (1-p^N)^n = 0$. Par le théorème des gendarmes, on peut donc conclure que $\lim\limits_{n\to+\infty} P(|M_n - N| > \varepsilon)$ existe et vaut 0. Donc (M_n) converge en probabilité vers N, soit en d'autres termes M_n est un estimateur convergent de N.

2. Les X_i admettent tous une même espérance et une même variance et sont indépendantes. Par la loi faible des grands nombres, $\overline{X_n}$ converge donc en probabilité vers $E(X_1) = Np$. $\overline{X_n}$ est donc un estimateur convergent de Np.

3. On a :

   - $M_n \xrightarrow{P} N$, donc $\dfrac{1}{M_n} \xrightarrow{P} \dfrac{1}{N}$.
   - $\overline{X_n} \xrightarrow{P} Np$, donc $\overline{X_n} \xrightarrow{\mathscr{L}} Np$ avec le résultat admis.

Par le théorème de Slutsky, on a donc que $\dfrac{\overline{X_n}}{M_n} \xrightarrow{\mathscr{L}} \dfrac{Np}{N} = p$. Par le résultat admis, on a donc $\dfrac{\overline{X_n}}{M_n} \xrightarrow{P} p$. Ainsi $\dfrac{\overline{X_n}}{M_n}$ est un estimateur convergent de p.

Exercice 19.13 (★★★ - Estimation du paramètre d'une loi de Poisson)

Soit $n \geq 2$ et $(X_1, \ldots, X_n)$ un échantillon de la loi de Poisson de paramètre θ inconnu (où $\theta \in]0, +\infty[$). On considère les variables :

$$U_n = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{[X_i=0]} \quad ; \quad S_n = \sum_{i=1}^{n} X_i \quad ; \quad T_n = \left(1 - \frac{1}{n}\right)^{S_n}.$$

1. Préciser la loi de S_n.
2. Montrer que U_n et T_n sont des estimateurs sans biais de $e^{-\theta}$.
3. Déterminer le risque quadratique $r_\theta(U_n)$ de U_n.
4. Montrer que le risque quadratique de T_n est :
$$r_\theta(T_n) = \exp(-2\theta)\left(\exp(\theta/n) - 1\right).$$
5. (a) Justifier que la fonction exponentielle est convexe. En déduire que :
$$e^{\theta/n} \leq \left(1 - \frac{1}{n}\right) + \frac{1}{n}e^{\theta}.$$
   (b) Comparer $r_\theta(T_n)$ et $r_\theta(U_n)$.

1. Par stabilité de la loi de Poisson par somme, les variables X_i étant indépendantes, S_n suit une loi $\mathscr{P}(n\theta)$.
2. Commençons par U_n. Rappelons que si $A \in \mathscr{A}$, alors la fonction indicatrice
$$\mathbb{1}_A : \omega \in \Omega \mapsto \begin{cases} 1 \text{ si } \omega \in A \\ 0 \text{ si } \omega \notin A \end{cases}$$
est une variable aléatoire réelle qui suit une loi de Bernoulli de paramètre $p = P(A)$.
Ainsi les variables $\mathbb{1}_{[X_i=0]}$ suivent une même loi de Bernoulli de paramètre $p = P(X_i = 0)$ (car les X_i suivent la même loi). De plus ces variables aléatoires sont indépendantes car les X_i le sont. Calculons plus précisément le paramètre :
$$p = P(X_1 = 0) = \frac{\theta^0}{0!}e^{-\theta} = e^{-\theta}.$$
Par linéarité de l'espérance, $E(U_n)$ existe et vaut :
$$E(U_n) = \frac{1}{n}\sum_{i=1}^{n} P(X_i = 0) = e^{-\theta}$$
Ainsi $b(U_n) = E(U_n) - e^{-\theta} = 0$ et U_n est un estimateur sans biais de $e^{-\theta}$.

Étudions l'estimateur T_n. Par théorème de transfert, $E(T_n)$ existe si et seulement si la série $\sum_{k\geq 0}\left(1-\frac{1}{n}\right)^k P(S_n=k)=\sum_{k\geq 0}\left(1-\frac{1}{n}\right)^k\frac{(n\theta)^k}{k!}e^{-n\theta}$ converge absolument, donc converge car la série est à termes positifs. Or on reconnait ici une série exponentielle de paramètre $x=\left(1-\frac{1}{n}\right)n\theta$. Elle converge donc bien. Ainsi $E(T_n)$ existe et on a :

$$E(T_n)=e^{\left(1-\frac{1}{n}\right)n\theta}e^{-n\theta}=e^{-\theta}.$$

En particulier $b(T_n)=E(T_n)-e^{-\theta}=0$ et T_n est un estimateur sans biais de $e^{-\theta}$.

3. On a par indépendance des $\mathbb{1}_{[X_i=0]}$:

$$V(U_n)=\frac{1}{n^2}\sum_{i=1}^n V(\mathbb{1}_{[X_i=0]})=\frac{1}{n^2}\sum_{i=1}^n V(\mathbb{1}_{[X_i=0]})=\frac{e^{-\theta}(1-e^{-\theta})}{n}.$$

D'où le risque quadratique :

$$r_\theta(U_n)=b_\theta(U_n)^2+V(U_n)=\frac{e^{-\theta}(1-e^{-\theta})}{n}.$$

4. On a :

$$T_n^2=\left(1-\frac{1}{n}\right)^{2S_n}.$$

Toujours par théorème de transfert, $E(T_n^2)$ existe si et seulement si la série $\sum_{k\geq 0}\left(1-\frac{1}{n}\right)^{2k}P(S_n=k)=\sum_{k\geq 0}\left(1-\frac{1}{n}\right)^{2k}\frac{(n\theta)^k}{k!}e^{-n\theta}$ converge absolument, donc converge car la série est à termes positifs. C'est encore une série exponentielle de paramètre $x=\left(1-\frac{1}{n}\right)^2 n\theta$. Elle converge donc bien et $E(T_n^2)$ existe et vaut :

$$E(T_n)=e^{\left(1-\frac{1}{n}\right)^2 n\theta}e^{-n\theta}=\exp\left(-2\theta+\frac{\theta}{n}\right).$$

Par la formule de Huygens, $V(T_n)$ existe et vaut :

$$V(T_n)=E(T_n^2)-E(T_n)^2=\exp\left(-2\theta+\frac{\theta}{n}\right)-e^{-2\theta}=e^{-2\theta}\left(e^{\frac{\theta}{n}}-1\right).$$

Enfin on a :

$$r_\theta(T_n)=b_\theta(T_n)^2+V(T_n)=V(T_n)=e^{-2\theta}\left(e^{\frac{\theta}{n}}-1\right).$$

5. (a) La fonction $f=\exp$ est de classe $\mathscr{C}^2$ sur $\mathbb{R}$ et on a $f''(x)=e^x\geq 0$ pour tout $x\in\mathbb{R}$. Donc f est convexe. Ainsi pour tout $x,y\in\mathbb{R}$ et pour tout $\lambda\in[0,1]$, on a :

$$e^{\lambda x+(1-\lambda)y}\leq \lambda e^x+(1-\lambda)e^y.$$

Prenons $x = \theta$, $y = 0$ et $\lambda = \dfrac{1}{n}$, on obtient :

$$e^{\theta/n} \leq \left(1 - \frac{1}{n}\right) e^0 + \frac{1}{n} e^\theta = \left(1 - \frac{1}{n}\right) + \frac{1}{n} e^\theta.$$

(b) On a avec l'inégalité précédente :

$$r_\theta(T_n) \leq e^{-2\theta} \left(\left(1 - \frac{1}{n}\right) + \frac{1}{n} e^\theta - 1\right) = e^{-2\theta} \frac{e^\theta - 1}{n} = r_\theta(U_n).$$

Ainsi T_n est un meilleur estimateur de θ que U_n.

Exercice 19.14 (★★★ - Oral ESCP 2012)

On cherche à évaluer le nombre N de lions d'Asie, espèce en voie de disparition, encore en vie dans la forêt de Gir. Pour cela, on capture d'abord en une seule fois m lions ($m \in \mathbb{N}^*$), que l'on tatoue avant de les relâcher dans la nature, et on admet que pendant toute la durée de l'étude, il n'y a ni naissance ni décès, puis l'on utilise l'une des deux méthodes suivantes.

1. **Méthode 1.** On capture successivement au hasard (donc avec équiprobabilité) et avec remise en liberté après observation du sujet, n lions. Soit Y_n le nombre de lions tatoués parmi eux.

   (a) Déterminer la loi de Y_n. En déduire que $\dfrac{1}{nm} Y_n$ est un estimateur sans biais et convergent de $\dfrac{1}{N}$.

   (b) Pourquoi ne peut-on pas prendre $\dfrac{nm}{Y_n}$ comme estimateur de N ?

   (c) On pose $B_n = \dfrac{m(n+1)}{Y_n + 1}$. Calculer l'espérance de B_n, et montrer que B_n est un estimateur asymptotiquement sans biais de N.

2. **Méthode 2.** On se donne $n \in \mathbb{N}^*$. On capture également, un par un, et avec remise en liberté après observation du sujet, n lions de Gir.

   On note X_n la variable aléatoire égale au nombre de lions qu'il a été juste nécessaire de capturer pour en obtenir n tatoués.

   On pose $D_1 = X_1$, et pour tout $i \in [\![2, n]\!]$, $D_i = X_i - X_{i-1}$. On admet que les D_i sont mutuellement indépendantes.

   (a) Pour tout $i \in [\![2, n]\!]$, que représente concrètement D_i ?

   (b) Déterminer, pour tout $i \in [\![2, n]\!]$, la loi de D_i, son espérance et sa variance. En déduire l'espérance et la variance de X_n.

   (c) On pose $A_n = \dfrac{m}{n} X_n$. Montrer que A_n est un estimateur sans biais et convergent de N, et déterminer son risque quadratique.
