Estimation

Feuille d'exercices

1 **Estimateurs sans biais de l'espérance et de la variance**
★ On considère une suite $(X_n)_{n\in\mathbb{N}^*}$ de variables aléatoires i.i.d. d'espérance m et de variance σ^2 avec $\sigma > 0$.

1. Montrer que la moyenne empirique
$$\overline{X}_n = \frac{1}{n}\sum_{k=1}^{n} X_k$$
est un estimateur sans biais et convergent de m.
2. On suppose dans cette question que m est connu. Montrer que
$$T_n = \frac{1}{n}\sum_{k=1}^{n}(X_k - m)^2$$
est un estimateur sans biais de σ^2.
3. On suppose dans cette question que m est inconnu.
   a. Montrer que
$$V_n = \frac{1}{n}\sum_{k=1}^{n}(X_k - \overline{X}_n)^2$$
est un estimateur asymptotiquement sans biais de σ^2 et calculer le biais de cet estimateur.
   b. En déduire un estimateur sans biais de σ^2.

La question de la convergence des estimateurs T_n et V_n est abordée dans l'exercice 7.

2 Soit $(X_1, \dots, X_n)$ un n-échantillon i.i.d. de loi uniforme sur $[0, \theta]$, où $\theta > 0$ est un paramètre inconnu. On pose
$$\overline{X}_n = \frac{1}{n}\sum_{k=1}^{n} X_k \quad \text{et} \quad M_n = \max(X_1, \dots, X_n).$$

1. Déterminer, en fonction de $\overline{X}_n$, un estimateur sans biais T_n de θ.
2. **a.** Justifier que M_n est une variable à densité et en donner une densité.
   b. Déterminer, en fonction de M_n, un estimateur sans biais U_n de θ.
3. Les estimateurs T_n et U_n de θ sont-ils convergents ?
4. Qui de T_n ou de U_n est le meilleur estimateur de θ lorsque $n \to \infty$?

3 1. **a.** Pour $(a, b) \in \mathbb{R} \times \mathbb{R}_+^*$, vérifier que la fonction
$$f_{a,b} : x \in \mathbb{R} \longmapsto \begin{cases} 0 & \text{si } x < a \\ \frac{1}{b}\exp\left(-\frac{x-a}{b}\right) & \text{si } x \geqslant a \end{cases}$$
est une densité de probabilité. On note $\mathcal{E}(a, b)$ la loi associée.
   b. Soit X une variable aléatoire de loi $\mathcal{E}(a, b)$. Reconnaître la loi de $\frac{X-a}{b}$ et en déduire les valeurs de $\mathbb{E}(X)$ et $\mathbb{V}(X)$.

On considère une suite $(X_n)_{n\in\mathbb{N}^*}$ de variables aléatoires i.i.d. de loi $\mathcal{E}(a, b)$ de paramètre $(a, b) \in \mathbb{R} \times \mathbb{R}_+^*$ inconnu. Pour tout $n \in \mathbb{N}^*$, on pose :
$$S_n = X_1 + \cdots + X_n, \quad T_n = \min(X_1, \dots, X_n) \quad \text{et} \quad U_n = \frac{S_n}{n} - T_n.$$

2. **a.** Montrer que T_n suit une loi $\mathcal{E}(\alpha, \beta)$ dont on déterminera les paramètres α et β.
   b. Montrer que T_n est un estimateur asymptotiquement sans biais et convergent de a.
3. **a.** Calculer $\mathbb{E}(U_n)$ pour $n \in \mathbb{N}^*$.
   b. Sans la calculer, justifier que $\operatorname{cov}(S_n, T_n)$ tend vers 0 lorsque $n \to \infty$.
   c. Justifier que U_n est un estimateur asymptotiquement sans biais et convergent de b.

4 Soit $(X_1, \dots, X_n)$ un n-échantillon i.i.d. de loi géométrique de paramètre $p \in]0, 1[$ inconnu. On pose $S_n = X_1 + \cdots + X_n$.

1. Montrer que $S_n(\Omega) = [\![n, +\infty[\![$ et :
$$\forall k \geqslant n, \quad \mathbb{P}(S_n = k) = \binom{k-1}{n-1}p^n q^{k-n}.$$
On dit que S_n suit une loi de Pascal de paramètre (n, p).
2. Montrer que $T_n = \frac{n-1}{S_n-1}$ est un estimateur sans biais de p.
3. En utilisant les théorèmes opératoires sur la convergence en probabilités vus en cours et en TD, montrer que T_n est un estimateur convergent de p.

5 On dispose de $n \geqslant 2$ observations indépendantes $X_1, \dots, X_n$ de même loi de Poisson de paramètre $\lambda > 0$ inconnu. On souhaite estimer le paramètre $\theta = \mathbf{e}^{-\lambda}$.
Pour tout $k \in [\![1, n]\!]$, on pose $Y_k = 1$ si $X_k = 0$ et $Y_k = 0$ sinon. On introduit également :
$$\overline{Y}_n = \frac{1}{n}\sum_{k=1}^{n} Y_k \quad \text{et} \quad S_n = \sum_{k=1}^{n} X_k.$$

1. **a.** Montrer que $\overline{Y}_n$ est un estimateur sans biais de θ.
   b. Calculer $\mathbb{V}(\overline{Y}_n)$ et en déduire que $\overline{Y}_n$ est un estimateur convergent de θ.
2. Pour $j \in \mathbb{N}$, calculer la probabilité conditionnelle $\varphi(j) = \mathbb{P}_{[S_n=j]}(X_1 = 0)$.
3. **a.** Montrer que $T_n = \varphi(S_n)$ est un estimateur sans biais de θ.
   b. Calculer $\mathbb{V}(T_n)$ et en déduire que T_n est un estimateur convergent de θ.
4. Comparer les risques quadratiques des deux estimateurs $\overline{Y}_n$ et T_n.

6 **Estimateur du maximum de vraisemblance**
★ 1. Pour $a > 0$, montrer que la fonction
$$f_a : x \longmapsto \begin{cases} 0 & \text{si } x < 1 \\ \frac{a}{x^{a+1}} & \text{si } x \geqslant 1 \end{cases}$$
est une densité de probabilité.

Dans la suite de l'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*

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

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 $h : a \longmapsto L(x_1, \dots, x_n, a)$ et montrer qu'elle atteint un maximum en un point $\widehat{a}$. On pourra considérer la fonction $g : a \longmapsto \ln\big(h(a)\big)$.

Puisque $\widehat{a}$ dépend de $x_1, \dots, x_n$, il existe une fonction $\varphi :]0, +\infty[^n \longrightarrow \mathbb{R}$ telle que $\widehat{a} = \varphi(x_1, \dots, x_n)$.
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 $S_n = Y_1 + \cdots + Y_n$.

4. Calculer $\mathbb{E}(T_n)$ et $\mathbb{V}(T_n)$.

5. En déduire que T_n est un estimateur asymptotiquement sans biais et convergent de a.

7 **Estimateur convergent de l'écart-type**

★ Soit X une variable aléatoire admettant un moment d'ordre 4, dont on note μ l'espérance et $\sigma > 0$ l'écart-type.
On considère un n-échantillon $X_1, \dots, X_n$ i.i.d. de même loi que X, dont on définit la moyenne et l'écart-type empiriques :

$$\overline{X}_n = \frac{1}{n}\sum_{k=1}^{n} X_k \quad \text{et} \quad S_n = \sqrt{\frac{1}{n}\sum_{k=1}^{n}\big(X_k - \overline{X}_n\big)^2}.$$

1. Montrer que :

$$S_n^2 = \frac{1}{n}\sum_{k=1}^{n} X_k^2 - \overline{X}_n^2.$$

2. En déduire que S_n^2 est un estimateur asymptotiquement sans biais de σ^2.

3. a. Que dire de la variable $\frac{1}{n}\sum_{k=1}^{n} X_k^2$ lorsque $n \to \infty$?
b. En utilisant les théorèmes opératoires sur la convergence en probabilité établis en travaux dirigés, en déduire que S_n^2 est un estimateur convergent de σ^2.
Remarque. Ce résultat peut également être établi à partir de la formule ci-dessous, dont la démonstration est assez technique :

$$\mathbb{V}(S_n^2) = \frac{(n-1)^2}{n^3}\mu_4 - \frac{(n-1)(n-3)}{n^3}\sigma^4$$

où $\mu_4 = \mathbb{E}\big((X-\mu)^4\big)$ désigne le moment centré d'ordre 4 de X.
c. Justifier que S_n est un estimateur convergent de σ.

8 ✎ On souhaite estimer la masse m d'un certain objet. Pour cela, on effectue des pesées successives et l'on note la moyenne obtenue. On admet que la variable aléatoire renvoyant le résultat d'une pesée de l'objet étudié suit la loi normale d'espérance m et d'écart-type $\sigma = 0{,}1$.

1. On effectue 10 pesées et on obtient une masse moyenne de 72,40 grammes. Donner un intervalle de confiance au niveau de confiance 90% pour la masse m.

2. Combien de pesées suffirait-il d'effectuer pour que la longueur de cet intervalle de confiance à 90% soit inférieure ou égale à 0,05 ?

3. Comment répondre à la question **1.** si l'on ne suppose plus que le résultat d'une pesée suit une loi normale ?

9 Soient $a \in]0, 1[$ et X une variable aléatoire à valeurs dans $[0, 1]$. On suppose que $\mathbb{P}(X \leqslant a) = \frac{1}{2}$, que conditionnellement à l'événement $[X \leqslant a]$, X suit la loi uniforme sur $[0, a]$ et que conditionnellement à l'événement $[X > a]$, elle suit la loi uniforme sur $[a, 1[$.

1. Déterminer la loi de X, son espérance et sa variance.

2. On cherche à estimer le paramètre inconnu a à partir d'un n-échantillon $(X_1, \dots, X_n)$ i.i.d. suivant la loi de X. On pose $\overline{X}_n = \frac{1}{n}\sum_{k=1}^{n} X_k$.
a. Déterminer en fonction de $\overline{X}_n$ un estimateur T_n sans biais de a.
b. L'estimateur T_n est-il convergent ?
c. Donner une majoration de $\mathbb{V}(T_n)$ en fonction de n uniquement puis en déduire les valeurs de $\varepsilon > 0$ pour lesquelles $[T_n - \varepsilon, T_n + \varepsilon]$ est un intervalle de confiance de a au niveau de risque $\alpha = 0{,}05$.
d. Proposer une deuxième méthode pour déterminer un intervalle de confiance asymptotique de a au niveau de confiance 95%.
e. Quelle est la meilleure méthode ?

10 Un institut de sondage désire estimer, parmi la population, la proportion p d'avis favorables à un projet donné. Afin de préserver la confidentialité des opinions individuelles, il réalise un sondage en respectant le protocole suivant.
Chaque personne interrogée doit, avant de répondre à la question posée, réaliser confidentiellement une expérience de Bernoulli de paramètre $\alpha \in]0, 1[\setminus \{\frac{1}{2}\}$ (elle seule en connaîtra le résultat). En cas de succès, elle répondra alors au sondeur selon ses convictions. En cas d'échec, elle répondra à l'opposé de ses convictions.
On note n le nombre de personnes interrogées et, pour tout $i \in [\![1, n]\!]$, X_i la variable aléatoire égale à 1 si la i-ième personne sondée exprime un avis favorable et à 0 sinon.

1. a. Montrer que les variables aléatoires $X_1, \dots, X_n$ suivent une même loi de Bernoulli dont on déterminera le paramètre q en fonction de α et p.
b. Proposer un estimateur S_n sans biais et convergent de q.
c. En déduire un estimateur T_n sans biais et convergent de p.
d. Calculer le risque quadratique de T_n. Quelles sont les contraintes qui pèsent sur le choix de α ? Comment choisir ce paramètre ?

2. a. On suppose que l'expérience de Bernoulli consiste à attendre un 6 lors d'un jet de dé. Sur 1000 personnes interrogées, le sondeur a recueilli 425 avis favorables. Donner un intervalle de confiance de p au niveau de confiance 95%.
   b. Écrire un script Scilab simulant le processus décrit dans la question **a.** et donnant 100 intervalles de confiance à 95%.

11 On cherche à évaluer le nombre N de poissons dans un étang. On prélève dans l'étang un échantillon de m poissons que l'on marque et que l'on remet dans l'étang. On propose deux méthodes différentes pour estimer N.

Première méthode

On prélève des poissons dans l'étang, au hasard et avec remise. Soit n un entier non nul inférieur ou égal à m.

Pour tout $i \in [\![1, n]\!]$, on note X_i le nombre de poissons qu'il a été necessaire de pêcher pour obtenir i poissons marqués.

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

1. a. Pour tout $i \in [\![1, n]\!]$, déterminer la loi de D_i, son espérance et sa variance.
   b. En déduire l'espérance et la variance de X_n.
   c. Montrer que $T_n = \frac{m}{n} X_n$ est un estimateur sans biais de N.
2. On suppose que $\sigma(T_n) \leqslant 100$. Déterminer un intervalle de confiance asymptotique de N au niveau de confiance 0,9. On donne $\Phi(1{,}64) \simeq 0{,}95$.
3. On a marqué 200 poissons puis effectué 450 prélèvements pour obtenir 50 poissons marqués. Déterminer la valeur observée de l'intervalle de confiance de N obtenu à la question **2.** pour l'échantillon étudié.

Deuxième méthode

On prélève successivement et avec remise n poissons. Soit Y_n le nombre de poissons marqués ainsi recueillis.

4. a. Montrer que $U_n = \frac{1}{nm} Y_n$ est un estimateur sans biais de $\frac{1}{N}$.
   b. Pour quelle raison ne peut-on pas considérer $\frac{nm}{Y_n}$ comme estimateur de N ?
5. On pose $V_n = \frac{m(n+1)}{Y_n+1}$.
   a. Calculer l'espérance de V_n.
   b. Quelle qualité de cet estimateur de N met-on ainsi en évidence ?

12 Soit $(X_n)_{n\geqslant 1}$ une suite de variables i.i.d. de loi de Poisson de paramètre $\lambda > 0$ inconnu. On pose

$$\overline{X}_n = \frac{1}{n} \sum_{k=1}^{n} X_k \quad \text{et} \quad T_n = \sqrt{n} \cdot \frac{\overline{X}_n - \lambda}{\sqrt{\lambda}}.$$

À l'aide de T_n, déterminer en fonction de $\overline{X}_n$ un intervalle de confiance asymptotique de λ au niveau de confiance $1 - \alpha$.

13 ♣ Soit $(X_n)_{n\geqslant 1}$ une suite de variables i.i.d. de loi géométrique de paramètre $p \in]0, 1[$ inconnu. On pose $q = 1 - p$ et, pour tout $n \in \mathbb{N}^*$,

$$\overline{X}_n = \frac{1}{n} \sum_{k=1}^{n} X_k, \quad Y_n = \frac{1}{\overline{X}_n} \quad \text{et} \quad T_n = \sqrt{n} \frac{p - Y_n}{Y_n \sqrt{1 - Y_n}}.$$

Montrer que :

$$\forall n \in \mathbb{N}^*, \quad T_n = \sqrt{\frac{q}{1 - Y_n}} \frac{\overline{X}_n - \frac{1}{p}}{\sqrt{\frac{q}{np^2}}}$$

et en déduire un intervalle de confiance asymptotique de p au niveau de confiance $1 - \alpha$.

Programmation :

- Classe : 1, 5, 6, 7, 10, 12
- TD : 2, 3, 8