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 $Z_n = \sum_{k=1}^{n} \frac{1+X_k}{2n}$ 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 $S_n = \frac{1}{n}\sum_{i=1}^{n} T_i^2$.

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 $S_n = \sum_{i=1}^{n} X_i$.

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

> 1. On a :
>
> $$S_n = \frac{1}{\lambda}\underbrace{(\lambda X_1 + \cdots + \lambda X_n)}_{=T_n}$$
>
> où les λX_i sont mutuellement indépendants (par lemme de coalition - les X_i le sont) et suivent une loi $\mathscr{E}(1) = \gamma(1)$. Par stabilité de la loi gamma, T_n suit une loi $\gamma(n)$. Notons $f_{T_n} : t \mapsto \begin{cases} \frac{1}{\Gamma(n)} t^{n-1} e^{-t} & \text{si } t > 0 \\ 0 & \text{sinon} \end{cases}$ une densité de T_n. Puisque $S_n = \frac{1}{\lambda} T_n$, S_n est une transformation affine d'une variable à densité. C'est donc une variable à densité également, et une densité de S_n est :
>
> $$f_{S_n} : t \mapsto \frac{1}{\left|\frac{1}{\lambda}\right|} f_{T_n}\left(\frac{t-0}{1/\lambda}\right) = \lambda f_{T_n}(\lambda t) = \begin{cases} \frac{\lambda}{\Gamma(n)} (\lambda t)^{n-1} e^{-\lambda t} & \text{si } t > 0 \\ 0 & \text{sinon} \end{cases}$$
>
> 2. Par le théorème de transfert, $Z_n = \frac{1}{S_n}$ admet une espérance si et seulement si l'intégrale $\int_{-\infty}^{+\infty} \frac{1}{t} f_{S_n}(t)\,\mathrm{d}t = \int_0^{+\infty} \frac{\lambda^n}{(n-1)!} t^{n-2} e^{-\lambda t}\,\mathrm{d}t$ converge absolument, donc converge puisque la fonction sous l'intégrale est positive. Par changement de variables affine (donc licite)

$u = \lambda t$, cette dernière intégrale est de même nature que :

$$\int_0^{+\infty} \frac{\lambda^n}{(n-1)!} \left(\frac{u}{\lambda}\right)^{n-2} e^{-u} \frac{\mathrm{d}u}{\lambda} = \frac{\lambda}{(n-1)!} \int_0^{+\infty} u^{n-2} e^{-u} \,\mathrm{d}u.$$

Or on reconnait ici une intégrale Gamma de paramètre $n-1$. Ainsi, $E(Z_n)$ existe bien et vaut :

$$E(Z_n) = \frac{\lambda}{(n-1)!}\Gamma(n-1) = \frac{\lambda}{n-1}.$$

En posant alors $U_n = (n-1)Z_n = \dfrac{n-1}{S_n}$, on a par linéarité de l'espérance que $E(U_n) = (n-1)\dfrac{\lambda}{n-1} = \lambda$, de sorte que U_n est 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) S_n^2 est-il un estimateur 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.
   (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 $(X_1, X_2, \dots, X_n)$ 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 $e^{-q\lambda}$? Que dire lorsque n est grand ? T_n est-il 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 (★★ - Risque quadratique - ✍)

Soit T_n un estimateur de $g(\theta)$. On suppose que pour tout $\theta \in \Theta$, T_n admet un moment d'ordre 2 pour la probabilité P_θ.
Pour tout $\theta \in \Theta$, on appelle *risque quadratique de T_n (en $g(\theta)$)* et on note $r_\theta(T_n)$ le réel :

$$r_\theta(T_n) = E_\theta\big[(T_n - g(\theta))^2\big].$$

Il représente la « dispersion moyenne » de T_n par rapport à $g(\theta)$.

1. Montrer que $r_\theta(T_n) = [E_\theta(T_n) - g(\theta)]^2 + V_\theta(T_n)$.
2. On suppose que pour tout $\theta \in \Theta$, $\lim\limits_{n\to+\infty} r_\theta(T_n) = 0$. Montrer que T_n est un estimateur convergent de $g(\theta)$.
3. Soient T_n et U_n deux estimateurs de $g(\theta)$ tels que $r_\theta(T_n) < r_\theta(U_n)$ pour tout $\theta \in \Theta$. Lequel de ces deux estimateurs doit-on privilégier ? Pourquoi ?
4. (a) **Exemple 1**. 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 $\overline{X_n}$ et T_n en tant qu'estimateurs de $\frac{1}{\lambda}$.
   (b) **Exemple 2.** On cherche à estimer le paramètre p d'une loi de Bernoulli. Soit $(X_1, \dots, X_n)$ un n-échantillon de loi mère $\mathscr{B}(p)$, et notons $S_n = \sum_{i=1}^n X_i$. On pose :

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

   Comparer $\overline{X_n}$ 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 ?

1.
2.
3.
4. (a)
   (b) Le calcul pour $\overline{X_n}$ a été fait de nombreuses fois, notamment en cours : $E(\overline{X_n}) = p$ et $V(\overline{X_n}) = \frac{V(X_1)}{n} = \frac{p(1-p)}{n}$.

Pour $T_n = \dfrac{S_n+1}{n+2}$, on a par linéarité de l'espérance que :

$$E(T_n) = \frac{E(S_n)+1}{n+2} = \frac{np+1}{n+2}.$$

Par propriété de la variance, on a :

$$V(T_n) = \frac{1}{(n+2)^2}V(S_n) = \frac{np(1-p)}{(n+2)^2}.$$

On en déduit que :

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

On a alors :

- pour p proche de $1/2$, $(E(\overline{X_n})-p)^2 + V(\overline{X_n}) \sim \dfrac{1}{4n}$ et $(E(T_n)-p)^2 + V(T_n) \sim \dfrac{n}{4(n+2)^2}$ de sorte que $(E(T_n)-p)^2 + V(T_n) < (E(\overline{X_n})-p)^2 + V(\overline{X_n})$;
- lorsque p tend vers 0, on a $(E(\overline{X_n})-p)^2 + V(\overline{X_n}) \to 0$ et $(E(T_n)-p)^2 + V(T_n) \sim \dfrac{1}{(n+2)^2}$, de sorte que $(E(T_n)-p)^2 + V(T_n) > (E(\overline{X_n})-p)^2 + V(\overline{X_n})$ quand p est proche de 0.

Le critère du cours ne nous permet donc pas de trancher entre ces deux estimateurs (on n'a pas $(E(T_n)-p)^2 + V(T_n) < (E(\overline{X_n})-p)^2 + V(\overline{X_n})$ pour tout $p \in]0,1[$ par exemple, ni l'inverse).

Exercice 19.8 (★★)

Soit θ un réel strictement positif, et soit $(X_k)_{k\in\mathbb{N}^*}$ une suite de variables i.i.d. suivant la loi uniforme sur $[0, 2\theta]$. On pose $M_n = \max(X_1, \dots, X_n)$ et $\overline{X_n} = \dfrac{1}{n}\sum_{i=1}^{n} X_i$.

1. Déterminer la loi de M_n, calculer son espérance et sa variance.
2. En déduire que $U_n = \dfrac{n+1}{2n}M_n$ est un estimateur sans biais de θ.
3. De $\overline{X_n}$ et U_n, lequel est un meilleur estimateur de θ ?

Exercice 19.9 (★★)

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

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

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

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

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

(b) i. Reconnaitre la loi de $Z_n = Y_n - \theta$.

ii. Y_n est-il un estimateur 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.

1. f_θ est une fonction positive et continue sauf éventuellement en θ. Considérons l'intégrale $I = \displaystyle\int_{-\infty}^{+\infty} f_\theta(t)\,\mathrm{d}t = \int_{\theta}^{+\infty} f_\theta(t)\,\mathrm{d}t$. On a pour tout $A > 0$:

$$\int_\theta^A f_\theta(t)\,\mathrm{d}t = \left[-e^{-(t-\theta)}\right]_\theta^A = 1 - e^{-(A-\theta)} \underset{n\to+\infty}{\longrightarrow} 1.$$

Ainsi I converge et vaut 1. f_θ est donc une densité de probabilité.

2. Posons $Y = X - \theta$. Par transformation affine d'une variable à densité, Y est à densité, de densité :

$$f_Y : t \in \mathbb{R} \mapsto \frac{1}{|1|} f_X\left(\frac{t-(-\theta)}{1}\right) = f_X(t+\theta) = \begin{cases} 0 & \text{si } t+\theta \leq \theta \\ e^{-((t+\theta)-\theta)} & \text{si } t+\theta > \theta \end{cases} = \begin{cases} 0 & \text{si } t \leq 0 \\ e^{-t} & \text{si } t+\theta > \theta \end{cases}.$$

On reconnait une densité de la loi exponentielle de paramètre 1, de sorte que $Y \hookrightarrow \mathscr{E}(1)$.

On a pour tout $x \in \mathbb{R}$:

$$\begin{aligned} F_X(x) &= P(X \leq x) = P(Y+\theta \leq x) = P(Y \leq x-\theta) \\ &= \begin{cases} 0 & \text{si } x-\theta \leq 0 \\ 1-e^{-(x-\theta)} & \text{si } x-\theta > 0 \end{cases} = \begin{cases} 0 & \text{si } x \leq \theta \\ 1-e^{-(x-\theta)} & \text{si } x > \theta \end{cases}. \end{aligned}$$

Par linéarité de l'espérance, $E(X)$ existe et on a :

$$E(X) = E(Y+\theta) = E(Y) + \theta = \frac{1}{1} + \theta = 1 + \theta.$$

D'autre part, $V(X)$ existe (par propriété de la variance) et vaut :

$$V(X) = V(Y+\theta) = V(Y) = \frac{1}{1^2} = 1.$$

3. (a) Par linéarité de l'espérance, $E(\overline{X_n})$ existe et vaut :

$$E(\overline{X_n}) = \frac{E(X_1) + \cdots + E(X_n)}{n} = \frac{n+n\theta}{n} = 1 + \theta.$$

On cherche alors T_n s'exprimant en fonction de $\overline{X_n}$ (et sans faire intervenir θ) tel que $E(T_n) = \theta$. On constate qu'en posant $T_n = \overline{X_n} - 1$, cela convient. Ainsi défini, T_n est donc un estimateur sans biais de θ.

Montrons que T_n est convergent. On peut calculer son risque quadratique et montrer qu'il tend vers 0. On peut aussi appliquer la loi faible des grands nombres comme suit : les X_k sont indépendantes et admettent toutes une même espérance $1+\theta$ et une même variance. Par la loi faible des grands nombres, on a :

$$\overline{X_n} \xrightarrow{P} 1 + \theta.$$

En composant par la fonction $f : x \mapsto x - 1$, on obtient :

$$T_n = f(\overline{X_n}) \xrightarrow{P} f(1+\theta) = \theta.$$

Ainsi T_n est un estimateur convergent de θ.

(b) i. Pour tout $x \in \mathbb{R}$, on a :

$$P(Z_n > x) = P(Y_n > x + \theta) = P(X_1 > x + \theta, \dots, X_n > x + \theta).$$

Les variables X_i étant indépendantes et suivant la même loi, on obtient :

$$P(Z_n > x) = P(X_1 > x + \theta)^n = (1 - P(X_1 \leq x + \theta))^n.$$

On obtient donc

$$\begin{aligned} P(Z_n \leq x) &= 1 - P(Z_n > x) = 1 - (1 - F_{X_1}(x+\theta))^n \\ &= \begin{cases} 1 - (1-0) & \text{si } x + \theta \leq \theta \\ 1 - (1 - (1 - e^{-((x+\theta)-\theta)})^n & \text{si } x + \theta > \theta \end{cases} = \begin{cases} 0 & \text{si } x \leq 0 \\ 1 - e^{-nx} & \text{si } x > 0 \end{cases}. \end{aligned}$$

Ainsi Z_n suit une loi $\mathscr{E}(n)$.

ii. Par linéarité de l'espérance, $E(Y_n)$ existe et vaut :

$$E(Y_n) = E(Z_n + \theta) = E(Z_n) + \theta = \frac{1}{n} + \theta \underset{n \to +\infty}{\longrightarrow} \theta.$$

Donc Y_n est un estimateur asymptotiquement sans biais de θ.

iii. On cherche ici aussi U_n qui s'exprime en fonction de Y_n (et sans faire intervenir θ) tel que $E(U_n) = \theta$. On voit que $U_n = Y_n - \frac{1}{n}$ convient. U_n est donc un estimateur sans biais de θ.

On a $U_n = Y_n - \frac{1}{n} = Z_n + \theta - \frac{1}{n}$, et $V(Z_n)$ existe et vaut $\frac{1}{n^2}$ puisque $Z_n \hookrightarrow \mathscr{E}(n)$. Par propriété de la variance, on en déduit que $V(U_n)$ existe et vaut :

$$V(U_n) = V\left(Z_n + \theta - \frac{1}{n}\right) = V(Z_n) = \frac{1}{n^2}.$$

On a :

$$E_\theta(U_n) = \theta \underset{n \to +\infty}{\longrightarrow} \theta \text{ et } V_\theta(U_n) = \frac{1}{n^2} \underset{n \to +\infty}{\longrightarrow} 0.$$

Donc U_n est un estimateur convergent de θ.

(c) On a $(E_\theta(U_n) - \theta)^2 + V_\theta(U_n) = \dfrac{1}{n^2}$.

Pour l'estimateur T_n, puisque les X_i admettent une variance qui vaut 1 et que ces variables sont indépendantes, on en déduit que $V(\overline{X_n})$ admet une variance, qui vaut :

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

Par propriété de la variance, $V(T_n)$ existe également et vaut :

$$V(T_n) = V(\overline{X_n} - 1) = V(\overline{X_n}) = \frac{1}{n}.$$

On a :
$$(E_\theta(T_n)-\theta)^2+V_\theta(T_n)=V_\theta(T_n)=\frac{1}{n}.$$
On constate que pour tout $\theta>0$, on a :
$$(E_\theta(U_n)-\theta)^2+V_\theta(U_n)<(E_\theta(T_n)-\theta)^2+V_\theta(T_n).$$
Donc U_n est un meilleur estimateur de θ que U_n (il renvoie en moyenne des meilleurs estimations du paramètre inconnu θ).

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 θ d'une loi de Poisson. Fixons pour cela `theta = 4` et $n=15$ pour commencer.
   (a) Créer un n-échantillon de la loi $\mathscr{P}(4)$ à l'aide de `Python`, et calculer les estimations obtenues de θ 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 $\mathscr{P}(\lambda)$,
   - calcule, pour chaque échantillon observé, l'estimation correspondante pour la moyenne empirique et pour la variance empirique (non biaisée),
   - trace l'histogramme des fréquences pour les 1000 estimations obtenues avec chacun des estimateurs (on pourra utiliser les commandes `plt.subplot(1,2,1)` et `plt.subplot(1,2,2)` pour tracer les histogrammes sur la même fenêtre graphique l'un à côté de l'autre).

   Comparer les deux histogrammes. Quel semble être le meilleur estimateur ?
   (c) Recommencer pour d'autres valeurs de θ et de n. Que constate-t-on en particulier lorsque n devient grand ?

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

1. Pour $a>0$, montrer que la fonction :
$$f_a : t\mapsto \begin{cases} 0 & \text{si } t\leq 1\\ \dfrac{a}{t^{a+1}} & \text{si } t>1\end{cases}$$
   est une densité de probabilité.

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

2. On définit la *fonction de vraisemblance*
$$L:(x_1,\ldots,x_n,a)\in\mathbb{R}^n\times\mathbb{R}_+^*\mapsto L(x_1,\ldots,x_n,a)=\prod_{k=1}^n f_a(x_k).$$

(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 $\varphi :]1, +\infty[^n \mapsto \mathbb{R}$ 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 $S_n = aY_1 + \cdots + aY_n$.
   (c) Calculer $E(T_n)$ et $V(T_n)$.
   (d) En déduire que T_n est un estimateur 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. 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. Puisque $M_n \xrightarrow{P} N$ et que $f : x \mapsto -\ln(x)$ est continue sur $\mathbb{R}_+^*$, $-\ln(M_n)$ converge en probabilité vers $-\ln(N)$. De même, $\ln(\overline{X_n})$ converge en probabilité vers $\ln(Np) = \ln(N) + \ln(p)$. Par somme, $\ln\left(\frac{\overline{X_n}}{M_n}\right) = \ln(\overline{X_n} - \ln(M_n)$ converge en probabilité vers $\ln(p)$. Et en passant à l'exponentielle qui est continue, on obtient bien la convergence en probabilité de $\frac{\overline{X_n}}{M_n}$ vers p.

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

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

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

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 $V_\theta(U_n)$.
4. Montrer que : $V_\theta(T_n) = \exp(-2\theta)\left(\exp(\theta/n) - 1\right).$
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 les estimateurs T_n et U_n de $e^{-\theta}$.

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 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 :

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

T_n est un estimateur sans biais de $e^{-\theta}$.

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

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

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 :

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

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 $\lambda=\frac{1}{n}$, 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) Les estimateurs T_n et U_n étant sans biais, il suffit de comparer leur variance. On a avec l'inégalité précédente :

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

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 relacher 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 $\frac{1}{nm}Y_n$ est un estimateur sans biais et convergent de $\frac{1}{N}$.

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

   (c) On pose $B_n = \frac{m(n+1)}{Y_n+1}$. Calculer l'espérence de B_n. Que dire de B_n comme estimateur de N lorsque n est grand ?

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 $A_n = \frac{m}{n}X_n$. Montrer que A_n est un estimateur sans biais et convergent de N.

1. (a) Il y a une proportion de $\frac{m}{N}$ lions tatoués. Y_n représente le nombre de succès dans une répétitions de n épreuves de Bernoulli de probabilité de succès $\frac{m}{N}$. Y_n suit donc une loi $\mathscr{B}(n, \frac{m}{N})$.

   Par linéarité de l'espérance $E\left(\frac{1}{nm}Y_n\right)$ existe et vaut $\frac{1}{nm}E(Y_n) = \frac{1}{nm}n\frac{m}{N} = \frac{1}{N}$. $\frac{1}{nm}Y_n$ est donc un estimateur sans biais de $\frac{1}{N}$.

   (b) Le fait que $\frac{Y_n}{mn}$ soit un estimateur sans biais de $\frac{1}{N}$ ne garantit pas en général que $\frac{mn}{Y_n}$ soit un estimateur sans biais de N.
   De manière plus terre à terre, $\frac{mn}{Y_n}$ a une probabilité non nulle de ne pas être défini : il suffit pour cela que Y_n soit nul, ce qui arrive avec une probabilité égale à $\left(1-\frac{m}{N}\right)^n \neq 0$.

   (c) Y_n, et donc B_n, étant une variable finie, $E(B_n)$ existe bien. Par le théorème de transfert, on a :

$$\begin{aligned}
E(B_n) &= \sum_{k=0}^{n} \frac{m(n+1)}{k+1} P(Y_n = k) = \sum_{k=0}^{n} \frac{m(n+1)}{k+1} \binom{n}{k} \left(\frac{m}{N}\right)^k \left(1 - \frac{m}{N}\right)^{n-k} \\
&= m \sum_{k=0}^{n} \binom{n+1}{k+1} \left(\frac{m}{N}\right)^k \left(1 - \frac{m}{N}\right)^{n-k} \\
&= N \sum_{k=0}^{n} \binom{n+1}{k+1} \left(\frac{m}{N}\right)^{k+1} \left(1 - \frac{m}{N}\right)^{n+1-(k+1)} \\
&= N \sum_{k=1}^{n+1} \binom{n+1}{k} \left(\frac{m}{N}\right)^k \left(1 - \frac{m}{N}\right)^{n+1-k} \\
&= N \left(\left(\sum_{k=0}^{n+1} \binom{n+1}{k} \left(\frac{m}{N}\right)^k \left(1 - \frac{m}{N}\right)^{n+1-k} \right) - \left(1 - \frac{m}{N}\right)^{n+1} \right) \\
&= N \left(\left(\frac{m}{N} + \left(1 - \frac{m}{N}\right) \right)^{n+1} - \left(1 - \frac{m}{N}\right)^{n+1} \right) \\
&= N \left(1 - \left(1 - \frac{m}{N}\right)^{n+1} \right)
\end{aligned}$$

B_n est donc un estimateur biaisé de N puisque $E(B_n) \neq N$. Cependant, on a

$$E(B_n) = N - N\left(1 - \frac{m}{N}\right)^{n+1} \underset{n\to+\infty}{\longrightarrow} N.$$

Plus n est grand, moins l'erreur commise en Donc B_n est un estimateur asymptotiquement sans biais de N.

2. (a) D_i représente le nombre de lions à capturer entre la capture du $(i-1)$-ème lion (exclu) et celle du i-ème lion (inclus).

   (b) Puisqu'on relâche les lions déjà capturés, à chaque capture de lion, la probabilité d'en avoir un qui est tatoué est $\frac{m}{N}$. Et donc D_i compte le nombre d'essais nécessaires à l'obtention d'un succès lors d'une répétition d'épreuves de Bernoulli indépendantes : D_i suit la loi géométrique de paramètre $\frac{m}{N}$. Par conséquent, on a :

$$E(D_i) = \frac{1}{\frac{m}{N}} = \frac{N}{m} \quad \text{et} \quad V(D_i) = \frac{1 - \frac{m}{N}}{\left(\frac{m}{N}\right)^2} = \frac{N(N-m)}{m^2}.$$

On a :

$$\begin{aligned}
X_n &= D_n + X_{n-1} = D_n + D_{n-1} + X_{n-2} = \cdots = D_n + D_{n-1} + \cdots + D_2 + X_1 \\
&= D_n + D_{n-1} + \cdots + D_2 + D_1.
\end{aligned}$$

Par linéarité de l'espérance, on a :

$$E(X_n) = \sum_{k=1}^{n} E(D_k) = \sum_{k=1}^{n} \frac{N}{m} = n\frac{N}{m}.$$

De plus, les D_k étant indépendants, on a :

$$V(X_n) = \sum_{k=1}^{n} V(D_k) = n\frac{N(N-m)}{m^2}.$$

(c) On a $E(A_n) = \dfrac{m}{n} E(X_n) = \dfrac{m}{n}\dfrac{nN}{m} = N$, donc A_n est un estimateur sans biais de N.
On a de plus :

$$V(A_n) = V(A_n) = \frac{m^2}{n^2} V(X_n) = \frac{m^2}{n^2} \frac{nN(N-m)}{m^2} = \frac{N(N-m)}{n} \underset{n \to +\infty}{\longrightarrow} 0.$$

A_n est donc un estimateur convergent de N.

Exercice 19.15 (★★★ - QSP HEC 2018)
Soit $(X_n)_{n \in \mathbb{N}^*}$ une suite de variables aléatoires indépendantes, suivant chacune la loi uniforme $\mathscr{U}([-\theta, \theta])$, où θ désigne un paramètre réel strictement positif inconnu.
Pour tout $n \in \mathbb{N}^*$, on note $U_n = \min(X_1, \dots, X_n)$ et $V_n = \max(X_1, \dots, X_n)$.

1. Démontrer que $(V_n)_{n \in \mathbb{N}^*}$ est une suite convergente d'estimateurs de θ.
2. Pour tout $n \in \mathbb{N}^*$, on note $T_n = \dfrac{1}{2} \sup \left\{X_j - X_i,\ (i,j) \in [\![1, n]\!]^2\right\}$.
   (a) Exprimer T_n en fonction de U_n et V_n.
   (b) En déduire que $(T_n)_{n \in \mathbb{N}^*}$ est une suite convergente d'estimateurs de θ.

1. Rappelons que la fonction de répartition des X_i est connue : il s'agit de

$$F_{X_i} : x \mapsto \begin{cases} 0 & \text{si } x \leqslant -\theta \\ \frac{x+\theta}{2\theta} & \text{si } -\theta \leqslant x \leqslant \theta \\ 1 & \text{si } x \geqslant \theta \end{cases}$$

Et donc, les X_i étant indépendantes, la fonction de répartition de V_n est

$$F_{V_n}(x) = \begin{cases} 0 & \text{si } x \leqslant -\theta \\ \left(\frac{x+\theta}{2\theta}\right)^n & \text{si } -\theta \leqslant x \leqslant \theta \\ 1 & \text{si } x \geqslant \theta \end{cases}$$

Pour x fixé, on a alors

$$\lim_{n \to +\infty} P(V_n \leqslant x) = \begin{cases} 0 & \text{si } x < \theta \\ 1 & \text{si } x \geqslant \theta \end{cases}$$

On reconnaît là la fonction de répartition d'une variable certaine égale à θ, vers laquelle (V_n) converge donc en loi.

La convergence en loi vers une constante est équivalente à la convergence en probabilité vers cette constante... mais ce résultat n'est malheureusement pas au programme, et notre définition de la convergence des estimateurs nécessite une convergence en probabilité.

Nous pouvons quand même nous en tirer sans grande difficulté : soit $\varepsilon > 0$. Alors

$$P(|V_n - \theta| \geq \varepsilon) = P(\theta - V_n \geq \varepsilon) = P(V_n \leq \theta + \varepsilon) = \begin{cases} 0 & \text{si } \varepsilon \geq \theta \\ (1 - \frac{\varepsilon}{2\theta})^n & \text{sinon} \end{cases} \underset{n \to +\infty}{\longrightarrow} 0.$$

Et donc $V_n \xrightarrow{P} \theta$, de sorte que V_n est bien un estimateur convergent de θ.

2. (a) On a $T_n = \frac{1}{2}(V_n - U_n)$: l'écart entre X_j et X_i va être maximal lorsque X_j est le plus grand possible et X_i le plus petit possible.

   (b) De la même manière qu'on a prouvé que $V_n \xrightarrow{P} \theta$, on prouverait que $U_n \xrightarrow{P} -\theta$.

   Pour la culture, notons tout de même que les $-X_i$ suivent la même loi que les X_i et que $U_n = -\max(-X_1, \ldots, -X_n)$.

   Ce maximum a la même loi que V_n et donc converge en probabilité vers θ, de sorte que U_n converge en probabilité vers $-\theta$.

   On a alors $T_n = \frac{1}{2}(V_n - U_n) \xrightarrow{P} \frac{1}{2}(\theta - (-\theta)) = \theta$ (la limite en probabilité d'une somme étant la somme des limites en probabilité).