Université Paris-Est Créteil **M2 IA2S**
2022–2023 **Probabilités**

Fiche n° 5 : Estimation Ponctuelle

Estimateurs

Exercice 1. Soit $(X_1, \dots, X_n)$ un n-échantillon de variables i.i.d. telles que $X_i \sim \mathcal{U}_{[0;\theta]}$ pour $i = 1, \dots, n$, où θ est inconnu.

1. Calculer l'espérance de X_i et en déduire un estimateur de θ.
2. Cet estimateur est-il sans biais ?
3. Cet estimateur est-il convergent ?
4. Cet estimateur est-il asymptotiquement normalement distribué ?

Corrigé 1.

1. On a $\mathbb{E}[X_i] = (\theta + 0)/2 = \theta/2$. Ainsi un estimateur de θ est $\widehat{\theta}_n = 2\overline{X}_n = \frac{2}{n}\sum_{i=1}^n X_i$.
2. Puisque les variables X_i sont i.i.d., on a
$$\mathbb{E}\left[\widehat{\theta}_n\right] = \frac{2}{n}\sum_{i=1}^n \mathbb{E}\left[X_i\right] = \frac{2}{n} \times \frac{n \times \theta}{2} = \theta.$$
Donc $\widehat{\theta}_n$ est sans biais.
3. On a $\mathbb{V}(X_i) = (\theta - 0)^2/12 = \theta^2/12$. Alors, comme les X_i sont i.i.d., on a
$$\mathbb{V}\left(\widehat{\theta}_n\right) = \frac{4}{n^2}\sum_{i=1}^n \mathbb{V}\left(X_i\right) = \frac{\theta^2}{3n} \underset{n\to+\infty}{\longrightarrow} 0.$$
Comme $\widehat{\theta}_n$, on obtient alors qu'il est convergent (au sens faible).
4. D'après le cours, on a que
$$\sqrt{n}\left(\overline{X}_n - \mathbb{E}[X_i]\right) \xrightarrow[n\to+\infty]{\text{loi}} \mathcal{N}\left(0; \mathbb{V}(X_i)\right),$$
ainsi
$$\sqrt{n}\left(\widehat{\theta}_n - \theta\right) = \sqrt{n}\left(2\overline{X}_n - 2\mathbb{E}[X_i]\right) \xrightarrow[n\to+\infty]{\text{loi}} \mathcal{N}\left(0; 4\mathbb{V}(X_i)\right) = \mathcal{N}\left(0; \frac{\theta^2}{3}\right).$$
Donc $\widehat{\theta}_n$ est asymptotiquement normal.

Exercice 2 (Comparaison d'estimateurs)**.** On considère une variable aléatoire continue X distribuée selon une loi de probabilité telle que $\mathbb{E}[X] = \theta$ et $\mathbb{V}(X) = \theta - \theta^2$, où θ est un paramètre inconnu vérifiant $\theta \in]0, 1[$. Soit un n-échantillon $(X_1, \dots, X_n)$ i.i.d. de même loi que X. Soient $\widehat{\theta}_n^1$ et $\widehat{\theta}_n^2$ deux estimateurs du paramètre θ respectivement définis par
$$\widehat{\theta}_n^1 = \frac{1}{n}\sum_{i=1}^n X_i, \quad \widehat{\theta}_n^2 = \frac{1}{n}\sum_{i=1}^n X_i^2.$$

1. Montrer que les estimateurs $\widehat{\theta}_n^1$ et $\widehat{\theta}_n^2$ sont sans biais.
2. Montrer que les estimateurs $\widehat{\theta}_n^1$ et $\widehat{\theta}_n^2$ sont convergents (au sens faible). On admettra que $\mathbb{V}(X^2) = 2\theta^2 - 2\theta^4$.
3. Peut-on déterminer quel est l'estimateur le plus précis ?
4. Les estimateurs $\widehat{\theta}_n^1$ et $\widehat{\theta}_n^2$ sont-ils asymptotiquement normalement distribués ?

Corrigé 2.

1. Calculons $\mathbb{E}\left[\widehat{\theta}_n^1\right]$ et $\mathbb{E}\left[\widehat{\theta}_n^2\right]$.

$$\mathbb{E}\left[\widehat{\theta}_n^1\right] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n X_i\right] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[X_i] = \frac{n \times \theta}{n} = \theta.$$

Par ailleurs, $\mathbb{E}[X_i^2] = \mathbb{V}(X_i) + \mathbb{E}[X_i]^2 = \theta - \theta^2 + \theta^2 = \theta$; d'où

$$\mathbb{E}\left[\widehat{\theta}_n^2\right] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n X_i^2\right] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[X_i^2] = \frac{n \times \theta}{n} = \theta.$$

Ainsi les deux estimateurs sont sans biais.

2. Calculons $\mathbb{V}\left(\widehat{\theta}_n^1\right)$ et $\mathbb{V}\left(\widehat{\theta}_n^2\right)$. Puisque les variables X_i sont i.i.d., on a

$$\mathbb{V}\left(\widehat{\theta}_n^1\right) = \frac{1}{n^2}\sum_{i=1}^n \mathbb{V}(X_i) = \frac{n \times (\theta - \theta^2)}{n^2} = \frac{\theta - \theta^2}{n},$$

$$\mathbb{V}\left(\widehat{\theta}_n^2\right) = \frac{1}{n^2}\sum_{i=1}^n \mathbb{V}(X_i^2) = \frac{n \times (2\theta^2 - 2\theta^4)}{n^2} = \frac{2\theta^2 - 2\theta^4}{n}.$$

Dans les deux cas, on a

$$\mathbb{E}\left[\widehat{\theta}_n^1\right] = \theta, \quad \mathbb{E}\left[\widehat{\theta}_n^2\right] = \theta,$$

$$\lim_{n\to+\infty} \mathbb{V}\left(\widehat{\theta}_n^1\right) = 0, \quad \lim_{n\to+\infty} \mathbb{V}\left(\widehat{\theta}_n^2\right) = 0.$$

Les deux estimateurs sont donc convergents au sens de la convergence en probabilité : $\widehat{\theta}_n^1 \xrightarrow[n\to+\infty]{\mathbb{P}} \theta$, $\widehat{\theta}_n^2 \xrightarrow[n\to+\infty]{\mathbb{P}} \theta$.

3. On sait que

$$\mathbb{V}\left(\widehat{\theta}_n^1\right) = \frac{\theta - \theta^2}{n}, \quad \mathbb{V}\left(\widehat{\theta}_n^2\right) = \frac{2\theta^2 - 2\theta^4}{n}.$$

Dès lors

$$\frac{\mathbb{V}\left(\widehat{\theta}_n^1\right)}{\mathbb{V}\left(\widehat{\theta}_n^2\right)} = \frac{\theta - \theta^2}{2\theta^2 - 2\theta^4} = \frac{\theta(1-\theta)}{2\theta^2(1-\theta^2)} = \frac{1}{2\theta(1+\theta)},$$

avec $\theta \in]0,1[$. Or sur $]0,1[$, la fonction $\theta \mapsto 2\theta(1+\theta)$ peut prendre des valeurs inférieures ou supérieures à 1, tout dépend donc de la vraie valeur du paramètre θ qui est inconnu. On ne peut donc pas déterminer quel est l'estimateur le plus précis en se basant sur la variance.

4. D'après le théorème central limite, on a

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}[X]\right) \xrightarrow[n\to+\infty]{\text{loi}} \mathcal{N}(0;\mathbb{V}(X)),$$

donc

$$\sqrt{n}\left(\widehat{\theta}_n^1 - \theta\right) \xrightarrow[n\to+\infty]{\text{loi}} \mathcal{N}(0;\theta-\theta^2).$$

De même,

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \mathbb{E}[X^2]\right) \xrightarrow[n\to+\infty]{\text{loi}} \mathcal{N}(0;\mathbb{V}(X^2)),$$

donc

$$\sqrt{n}\left(\widehat{\theta}_n^2 - \theta\right) \xrightarrow[n\to+\infty]{\text{loi}} \mathcal{N}(0;2\theta^2-2\theta^4).$$

MAXIMUM DE VRAISEMBLANCE

Exercice 3. On considère un n-échantillon $(X_1,\ldots,X_n)$ de variables aléatoires i.i.d. de distribution de Bernoulli $\mathcal{B}(p)$, où $p \in [0;1]$ est un paramètre inconnu. On dispose de réalisations $(x_1,\ldots,x_n)$, où $x_i \in \{0,1\}$, $i=1,\ldots,n$.

1. Donner la fonction de masse de la variable aléatoire X_i. En déduire la vraisemblance de l'échantillon ainsi que la log-vraisemblance.
2. Écrire les conditions du premier et second ordres. En déduire l'estimateur du maximum de vraisemblance de p.
3. L'estimateur du maximum de vraisemblance de p est-il efficace ?

Corrigé 3.

1. La fonction de masse de la variable aléatoire X_i s'écrit

$$f_X(x_i;p) = \mathbb{P}(X_i = x_i;p) = p^{x_i}(1-p)^{1-x_i}, \quad x_i \in \{0,1\}.$$

On en déduit la fonction de vraisemblance de l'échantillon

$$L(x_1,\ldots,x_n;p) = \prod_{i=1}^{n}\mathbb{P}(X_i = x_i;p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i},$$

et la log-vraisemblance de l'échantillon

$$\ln L(x_1,\ldots,x_n;p) = \ln p\sum_{i=1}^{n} x_i + \ln(1-p)\left(n-\sum_{i=1}^{n} x_i\right).$$

2. La condition du premier ordre du programme de maximisation de la log-vraisemblance s'écrit

$$\begin{aligned}
\text{C1 :} \quad & \left.\frac{\partial \ln L(x_1,\ldots,x_n;p)}{\partial p}\right|_{p=\widehat{p}_n} = 0 \\
\Longleftrightarrow \quad & \frac{1}{\widehat{p}_n}\sum_{i=1}^{n} x_i - \frac{1}{1-\widehat{p}_n}\left(n-\sum_{i=1}^{n} x_i\right) = 0 \Longleftrightarrow (1-\widehat{p}_n)\sum_{i=1}^{n} x_i - \widehat{p}_n\left(n-\sum_{i=1}^{n} x_i\right) = 0 \\
\Longleftrightarrow \quad & \sum_{i=1}^{n} x_i - n\widehat{p}_n = 0 \Longleftrightarrow \widehat{p}_n(x) = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}_n \Longleftrightarrow \widehat{p}_n^{MV} = \overline{X}_n.
\end{aligned}$$

On vérifie maintenant la condition du second ordre du programme de maximisation de la log-vraisem-blance pour $\widehat{p}_n(x) = \bar{x}_n$.

$$\text{C2 :}\frac{\partial^2 \ln L(x_1,\ldots,x_n;p)}{\partial p^2}\bigg|_{p=\bar{x}_n} = -\frac{\sum_{i=1}^n x_i}{\bar{x}_n^2} - \frac{n-\sum_{i=1}^n x_i}{(1-\bar{x}_n)^2} = -\frac{n\bar{x}_n}{\bar{x}_n^2} - \frac{n-n\bar{x}_n}{(1-\bar{x}_n)^2} = -\frac{n}{\bar{x}_n} - \frac{n}{(1-\bar{x}_n)} < 0$$

car $0 < \bar{x}_n < 1$. Nous avons donc bien un maximum. Par conséquent, l'estimateur du maximum de vraisemblance du paramètre p correspond à la moyenne empirique.

3. Le score s'écrit

$$\begin{aligned} S_n(X,p) &= \frac{\partial \ln L(X_1,\ldots,X_n;p)}{\partial p} = \frac{1}{p}\sum_{i=1}^n X_i - \frac{1}{1-p}\left(n - \sum_{i=1}^n X_i\right) \\ &= \frac{n}{p}\overline{X}_n - \frac{n}{1-p}\left(1-\overline{X}_n\right) = \frac{n}{p(1-p)}\overline{X}_n - \frac{n}{1-p}. \end{aligned}$$

Comme $\mathbb{V}(\overline{X}_n) = \dfrac{p(1-p)}{n}$, on a

$$I_n(p) = \mathbb{V}(S_n(X,p)) = \left(\frac{n}{p(1-p)}\right)^2 \mathbb{V}\left(\overline{X}_n\right) = \frac{n}{p(1-p)}.$$

Ainsi $\overline{X}_n$ est un estimateur efficace de p puisque $\mathbb{V}(\overline{X}_n) = I_n(p)^{-1}$.

Exercice 4. On considère un n-échantillon $(X_1,\ldots,X_n)$ de variables aléatoires i.i.d. de distribution géométrique $\mathcal{G}(p)$, où $p \in [0;1]$ est un paramètre inconnu. On dispose de réalisations $(x_1,\ldots,x_n)$, où $x_i \in \mathbb{N}^*$, $i = 1,\ldots,n$.

1. Donner la fonction de masse de la variable aléatoire X_i. En déduire la vraisemblance de l'échantillon ainsi que la log-vraisemblance.
2. Écrire les conditions du premier et second ordres. En déduire l'estimateur du maximum de vraisemblance de p.

Corrigé 4. 1. La fonction de masse de la variable aléatoire X_i s'écrit

$$f_X(x_i;p) = \mathbb{P}(X_i = x_i;p) = p(1-p)^{x_i-1}, \quad x_i \in \mathbb{N}^*.$$

On en déduit la fonction de vraisemblance de l'échantillon

$$L(x_1,\ldots,x_n;p) = \prod_{i=1}^n \mathbb{P}(X_i = x_i;p) = \prod_{i=1}^n p(1-p)^{x_i-1} = p^n(1-p)^{\sum_{i=1}^n x_i - n},$$

et la log-vraisemblance de l'échantillon

$$\ln L(x_1,\ldots,x_n;p) = n\ln p + \ln(1-p)\left(\sum_{i=1}^n x_i - n\right).$$

2. La condition du premier ordre du programme de maximisation de la log-vraisemblance s'écrit

$$\text{C1 :} \quad \left.\frac{\partial \ln L(x_1,\ldots,x_n;p)}{\partial p}\right|_{p=\widehat{p}_n} = 0$$

$$\iff \frac{n}{\widehat{p}_n} - \frac{1}{1-\widehat{p}_n}\left(\sum_{i=1}^n x_i - n\right) = 0 \iff n(1-\widehat{p}_n) - \widehat{p}_n\left(\sum_{i=1}^n x_i - n\right) = 0$$

$$\iff n - \widehat{p}_n \sum_{i=1}^n x_i = 0 \iff \widehat{p}_n(x) = \frac{n}{\sum_{i=1}^n x_i} = \frac{1}{\bar{x}_n} \iff \widehat{p}_n^{MV} = \frac{1}{\overline{X}_n}.$$

On vérifie maintenant la condition du second ordre du programme de maximisation de la log-vraisem-blance pour $\widehat{p}_n(x) = 1/\bar{x}_n$.

$$\begin{aligned}\text{C2 :} \quad \left.\frac{\partial^2 \ln L(x_1,\ldots,x_n;p)}{\partial p^2}\right|_{p=1/\bar{x}_n} &= \left.-\frac{n}{p^2} - \frac{\sum_{i=1}^n x_i - n}{(1-p)^2}\right|_{p=1/\bar{x}_n} = -n\bar{x}_n^2 - \frac{n\bar{x}_n - n}{(1-1/\bar{x}_n)^2} \\ &= -n\bar{x}_n^2 - \frac{n(\bar{x}_n-1)\bar{x}_n^2}{(\bar{x}_n-1)^2} = -n\bar{x}_n^2\left(1+\frac{1}{\bar{x}_n-1}\right) \\ &= -\frac{n\bar{x}_n^3}{\bar{x}_n-1} < 0\end{aligned}$$

car $\bar{x}_n > 1$. Nous avons donc bien un maximum. Par conséquent, l'estimateur du maximum de vraisemblance du paramètre p correspond à l'inverse de la moyenne empirique.

Exercice 5. On considère un n-échantillon $(X_1,\ldots,X_n)$ de variables aléatoires i.i.d. telles que la densité de la variable aléatoire X_i s'écrive

$$f_X(x_i;\theta) = \frac{x_i}{\theta}\exp\left(-\frac{x_i^2}{2\theta}\right), \quad x_i \in \mathbb{R}_+$$

où $\theta > 0$ est un paramètre inconnu. On dispose de réalisations $(x_1,\ldots,x_n)$, où $x_i \in \mathbb{R}_+$, $i = 1,\ldots,n$.

1. Écrire la vraisemblance de l'échantillon ainsi que la log-vraisemblance.
2. Écrire les conditions du premier et second ordres. En déduire l'estimateur du maximum de vraisemblance de θ.
3. Sachant que, pour tout $1 \le i \le n$, $\mathbb{E}[X_i] = \sqrt{\pi\theta/2}$, $\mathbb{E}[X_i^2] = 2\theta$ et $\mathbb{E}[X_i^4] = 8\theta^2$, vérifier que l'estimateur du maximum de vraisemblance de θ est sans biais, convergent et efficace.

Corrigé 5. 1. La fonction de vraisemblance de l'échantillon s'écrit

$$L(x_1,\ldots,x_n;\theta) = \prod_{i=1}^n f_X(x_i;\theta) = \prod_{i=1}^n \frac{x_i}{\theta}\exp\left(-\frac{x_i^2}{2\theta}\right) = \frac{1}{\theta^n}\prod_{i=1}^n x_i \exp\left(-\frac{1}{2\theta}\sum_{i=1}^n x_i^2\right),$$

et la log-vraisemblance de l'échantillon

$$\ln L(x_1,\ldots,x_n;\theta) = -n\ln(\theta) + \sum_{i=1}^n \ln(x_i) - \frac{1}{2\theta}\sum_{i=1}^n x_i^2.$$

2. La condition du premier ordre du programme de maximisation de la log-vraisemblance s'écrit

$$
\begin{aligned}
\text{C1 :} \qquad & \left.\frac{\partial \ln L(x_1,\ldots,x_n;\theta)}{\partial \theta}\right|_{\theta=\widehat{\theta}_n} = 0 \\
\iff \quad & -\frac{n}{\widehat{\theta}_n} + \frac{1}{2(\widehat{\theta}_n)^2}\sum_{i=1}^n x_i^2 = 0 \iff \sum_{i=1}^n x_i^2 = 2n\widehat{\theta}_n \iff \widehat{\theta}_n(x) = \frac{1}{2n}\sum_{i=1}^n x_i^2 \\
\iff \quad & \widehat{\theta}_n^{MV} = \frac{1}{2n}\sum_{i=1}^n X_i^2.
\end{aligned}
$$

On vérifie maintenant la condition du second ordre du programme de maximisation de la log-vraisem-blance pour $\widehat{\theta}_n(x) = \frac{1}{2n}\sum_{i=1}^n x_i^2$.

$$
\text{C2 :} \quad \left.\frac{\partial^2 \ln L(x_1,\ldots,x_n;\theta)}{\partial \theta^2}\right|_{\theta=\widehat{\theta}_n} = \frac{n}{(\widehat{\theta}_n)^2} - \frac{1}{(\widehat{\theta}_n)^3}\sum_{i=1}^n x_i^2 = \frac{n}{(\widehat{\theta}_n)^2} - \frac{2n\widehat{\theta}_n}{(\widehat{\theta}_n)^3} = -\frac{n}{(\widehat{\theta}_n)^2} < 0.
$$

Nous avons donc bien un maximum. Par conséquent, l'estimateur du maximum de vraisemblance du paramètre θ est $\widehat{\theta}_n^{MV} = \frac{1}{2n}\sum_{i=1}^n X_i^2$.

3. On a

$$
\mathbb{E}\left[\widehat{\theta}_n^{MV}\right] = \frac{1}{2n}\sum_{i=1}^n \mathbb{E}[X_i^2] = \frac{1}{2n}\sum_{i=1}^n (2\theta) = \theta,
$$

donc $\widehat{\theta}_n^{MV}$ est un estimateur sans biais de θ. De plus, comme les X_i sont indépendants, on a

$$
\begin{aligned}
\mathbb{V}\left(\widehat{\theta}_n^{MV}\right) &= \frac{1}{4n^2}\sum_{i=1}^n \mathbb{V}(X_i^2) = \frac{1}{4n^2}\sum_{i=1}^n \left(\mathbb{E}[X_i^4] - \mathbb{E}[X_i^2]^2\right) = \frac{1}{4n^2}\sum_{i=1}^n \left(8\theta^2 - (2\theta)^2\right) \\
&= \frac{1}{4n^2}\sum_{i=1}^n 4\theta^2 = \frac{\theta^2}{n} \underset{n\to+\infty}{\longrightarrow} 0,
\end{aligned}
$$

donc $\widehat{\theta}_n^{MV}$ est un estimateur convergent de θ. Calculons le score pour vérifier qu'il est efficace. On a

$$
S_n(X,\theta) = \frac{\partial \ln L(X_1,\ldots,X_n;\theta)}{\partial \theta} = -\frac{n}{\theta} + \frac{1}{2\theta^2}\sum_{i=1}^n X_i^2 = -\frac{n}{\theta} + \frac{n}{\theta^2}\widehat{\theta}_n^{MV}.
$$

Comme $\mathbb{V}(\widehat{\theta}_n^{MV}) = \dfrac{\theta^2}{n}$, on a

$$
I_n(\theta) = \mathbb{V}(S_n(X,\theta)) = \left(\frac{n}{\theta^2}\right)^2 \mathbb{V}\left(\widehat{\theta}_n^{MV}\right) = \frac{n}{\theta^2}.
$$

Ainsi $\widehat{\theta}_n^{MV}$ est un estimateur efficace de θ puisque $\mathbb{V}\left(\widehat{\theta}_n^{MV}\right) = I_n(\theta)^{-1} = \dfrac{\theta^2}{n}$.