Chapitre 19

Estimation ponctuelle

Compétences attendues.

- ✓ Montrer qu'un estimateur est sans biais.
- ✓ Montrer qu'un estimateur est convergent.
- ✓ Comparer deux estimateurs.

Mathieu Mansuy

Professeur en ECG deuxième année spécialité mathématiques approfondies au Lycée Louis Pergaud (Besançon)

Page personnelle : `mathieu-mansuy.fr/`

E-mail : `mathieu.mansuy@ac-besancon.fr`

1 Position du problème

On étudie un phénomène aléatoire qui est reproductible dans des conditions identiques et indépendantes. On connait pour des raisons théoriques ou empiriques le type de loi le décrivant. Mais les paramètres de la dite loi sont souvent inconnus. On doit donc les estimer : c'est l'objectif de ce qu'on appelle la statistique inférentielle.

Exemples.

- À l'approche du second tour d'une élection présidentielle, on interroge une personne au hasard et on note $X = 1$ si elle se prononce pour le candidat A et $X = 0$ si c'est pour le candidat B. X suit une loi de Bernoulli de paramètre $\theta_0 \in [0, 1]$ inconnu qui correspond à la proportion de français qui votent pour A.
- On souhaite modéliser le nombre N de voitures se présentant à un péage en une heure. Il s'agit du nombre de réalisations d'un évènement rare sur un grand nombre d'observations. On sait donc que ce nombre N suit une loi de Poisson. On cherche le paramètre de cette loi.
- On souhaite modéliser la durée de vie T d'un appareil électrique. Ce phénomène étant sans mémoire, il se modélise à l'aide d'une loi exponentielle. Il s'agit de trouver son paramètre.

Notons X la variable égale au résultat de notre expérience aléatoire. On suppose donc que l'on ne connaît qu'imparfaitement la loi de X : on sait de quel type elle est (elle appartient à une famille de lois $(\mu_\theta)_{\theta\in\Theta}$ connue), mais elle dépend d'un paramètre θ inconnu appartenant à Θ, l'espace des paramètres. Le paramètre θ peut être réel (Θ est une partie $\mathbb{R}$) ou vectoriel (Θ est une partie de $\mathbb{R}^k$, $k \geq 2$).

Exemple. Dans les trois cas précédents :

- $X \hookrightarrow \mathscr{B}(\theta)$, $\Theta =]0, 1[$;
- $N \hookrightarrow \mathscr{P}(\theta)$, $\Theta =]0, +\infty[$;
- $T \hookrightarrow \mathscr{E}(\theta)$, $\Theta =]0, +\infty[$.

On peut aussi imaginer le cas de $X \hookrightarrow \mathscr{N}(m, \sigma^2)$ où $\theta = (m, \sigma^2)$ est inconnu (paramètre vectoriel, $\Theta = \mathbb{R}\times]0, +\infty[$).

L'objectif de la statistique inférentielle est d'estimer la vraie valeur[1] θ_0 du paramètre θ à partir de réalisations[2] de la variable X.

2 Vocabulaire de l'estimation

2.1 Échantillonage

Bien entendu, une seule réalisation d'une variable aléatoire de loi μ ne permettra pas d'obtenir beaucoup d'informations sur μ. On est donc amené à introduire la notion d'échantillon.

Définition.

- On appelle *n-échantillon de loi mère μ* sur un espace probabilisé $(\Omega, \mathscr{A}, P)$ un n-uplet $(X_1, \ldots, X_n)$ de variables aléatoires i.i.d. suivant toutes la loi μ.
- Un n-uplet $(x_1, \ldots, x_n) \in \mathbb{R}^n$ est une *réalisation de l'échantillon* (ou *échantillon observé*) si c'est une réalisation du vecteur aléatoire $(X_1, \ldots, X_n)$, c'est-à-dire si :

$$\exists \omega \in \Omega, \quad (x_1, \ldots, x_n) = (X_1(\omega), \ldots, X_n(\omega)).$$

[1]En pratique, on désignera la variable θ et la vraie valeur θ_0 que l'on cherche à évaluer par la même lettre θ.
[2]Par réalisation de X, on entend la valeur $x = X(\omega)$ que prend la variable X en une certaine éventualité ω.

Remarque. Un échantillon est un n-uplet de variables aléatoires alors qu'un échantillon observé est un n-uplet de réels. Si par exemple $\mathscr{E}(\theta)$ est la loi mère du n-échantillon, alors `rd.exponential(1/theta,n)` fournit une réalisation de cet échantillon.

Pour aller plus loin. Modèle statistique.

Soit $(x_1, \dots, x_n) \in \mathbb{R}^n$ un échantillon de données obtenu en observant n fois le phénomène aléatoire. On admettra l'existence de n variables aléatoires $X_1, \dots, X_n$, toutes définies sur un même espace probabilisable $(\Omega, \mathscr{A})$, telles que $(x_1, \dots, x_n)$ soit une réalisation de $(X_1, \dots, X_n)$, c'est-à-dire :

$$\exists\, \omega \in \Omega, \quad (X_1(\omega), \dots, X_n(\omega)) = (x_1, \dots, x_n).$$

Pour tout $\theta \in \Theta$, on admettra de plus l'existence d'une probabilité P_θ sur $(\Omega, \mathscr{A})$ telle que $(X_1, \dots, X_n)$ soit un n-échantillon P_θ-indépendant de loi mère μ_θ.

Remarque. Ne connaissant pas la valeur de θ_0, ce modèle statistique va nous permettre d'effectuer nos calculs sous toutes les probabilités P_θ, et ainsi d'obtenir des résultats valables quelque soit sa valeur réelle.

Notation.

Si X est une variable aléatoire sur $(\Omega, \mathscr{A})$, on notera, en cas d'existence, $E_\theta(X)$ et $V_\theta(X)$ son espérance et sa variance pour la probabilité P_θ.

Exemple. Reprenons l'exemple de l'élection présidentielle. On questionne $n = 5$ individus sur leurs intentions de vote et on obtient les résultats suivants[3] (en notant 1 ou 0 selon que le choix se porte sur le candidat A ou B) :

$$1, \quad 0, \quad 0, \quad 1, \quad 0.$$

Ces résultats observés correspondent, pour tout $\theta \in [0, 1]$, à la réalisation d'un 5-échantillon $(X_1, \dots, X_5)$ P_θ-indépendant de loi mère $\mathscr{B}(\theta)$. Pour tout $1 \leq i \leq 5$, on a $E_\theta(X_i) = \theta$ et $V_\theta(X_i) = \theta(1 - \theta)$.

2.2 Estimateur

À partir d'un échantillon observé, on souhaite estimer une valeur caractéristique de la loi μ_θ telle que son espérance, sa variance, son étendue... On notera $g(\theta)$ cette valeur, où $g : \Theta \to \mathbb{R}$.

Définition.

- On appelle *estimateur* (d'ordre n) de $g(\theta)$ toute variable aléatoire T_n de la forme $\varphi(X_1, \dots, X_n)$, où $(X_1, \dots, X_n)$ est un n-échantillon.
- On appelle *estimation* de $g(\theta)$ une réalisation $t_n = \varphi(x_1, \dots, x_n)$ de $T_n = \varphi(X_1, \dots, X_n)$.

Remarques.

- Un estimateur de $g(\theta)$ est une variable aléatoire alors qu'une estimation de $g(\theta)$ est un réel.
- L'estimateur T_n ne dépend que du n-échantillon $(X_1, \dots, X_n)$. En particulier, φ ne doit pas dépendre du paramètre θ qui est inconnu et qu'on cherche justement à estimer.

[3] En toute rigueur, on ne sonde pas deux fois la même personne, et les résultats obtenus ne correspondent donc pas à une même expérience reproduite de manière identique et indépendante. Mais n étant négligeable devant la population française, on peut raisonnablement le supposer.

Définition.

On appelle *moyenne empirique* l'estimateur :

$$\overline{X_n} = \frac{X_1 + \cdots + X_n}{n}.$$

Exemple. Reprenons l'exemple de l'élection présidentielle. La moyenne empirique $\overline{X_n} = \frac{1}{n}\sum_{k=1}^{n} X_k$ est l'estimateur « le plus naturel » de θ. On obtient avec l'échantillon observé que :

$$\overline{X_5}(\omega) = \frac{1}{5}(1+0+0+1+0) = \frac{2}{5} \text{ est une estimation de } \theta.$$

On peut envisager bien d'autres estimateurs pour θ, comme par exemple :

$$A_n = \frac{2}{n(n+1)}\sum_{k=1}^{n} kX_k \quad \text{ou} \quad B_n = 0,$$

chacun fournissant une autre estimation de θ à partir de notre échantillon observé :

$$A_n(\omega) = \frac{2}{5 \times 6}(1+0+0+4+0) = \frac{1}{3} \quad \text{ou} \quad B_n(\omega) = 0.$$

On verra dans la suite comment décider :

- s'il est pertinent de prendre ces nombres comme estimation de θ ou non[4] ;
- si l'un de ces estimateurs est plus pertinent que les autres, c'est-à-dire s'il donne en moyenne des estimations « plus proches » de la véritable valeur de θ.

Exemple. Si $(X_1, \ldots, X_n)$ est un n-échantillon de loi mère $\mathscr{E}(\theta)$, alors $\overline{X_n}$ est un estimateur de $g(\theta) = \frac{1}{\theta}$, et `np.mean(rd.exponential(1/theta,n))` en est une estimation.

Exemple. Si $(X_1, \ldots, X_n)$ est un n-échantillon de loi mère $\mathscr{U}([a,b])$ de paramètre vectoriel inconnu $\theta = (a,b) \in \mathbb{R}^2$, alors :

$$T_n = \max(X_1, \ldots, X_n) - \min(X_1, \ldots, X_n)$$

est un estimateur de $g(\theta) = b - a$. Beaucoup d'autres estimateurs sont possibles, comme le suivant[5] :

$$U_n = 2\overline{X_n}.$$

Remarque. Très souvent, on prendra $g = \mathrm{Id}_{\mathbb{R}}$ et on estimera θ directement. Mais on peut aussi chercher à estimer une fonction $g(\theta)$ des paramètres, comme dans l'exemple précédent pour l'étendue d'une loi uniforme. On peut aussi vouloir estimer $g(\theta)$, et non θ directement, par facilité de calculs : pour la loi exponentielle de paramètre θ par exemple, il est plus facile d'estimer $g(\theta) = \frac{1}{\theta}$ en utilisant la moyenne empirique.

[4] Bien sûr, l'estimation de θ obtenue par B_n n'est absolument pas pertinente, même si B_n répond bien lui aussi à la définition d'estimateur.

[5] Pourquoi cet estimateur est-il pertinent ?

3 Comparaison d'estimateurs

3.1 Estimateur sans biais

Définition.

Soit T_n un estimateur de $g(\theta)$. On suppose que pour tout $\theta \in \Theta$, T_n admet une espérance $E_\theta(T_n)$ pour la probabilité P_θ.

On dit que T_n est un *estimateur sans biais de* $g(\theta)$ lorsque :

$$\forall \theta \in \Theta, \quad E_\theta(T_n) = g(\theta).$$

Dans le cas contraire, on dit que T_n est un estimateur *biaisé*.

Remarque. Un estimateur T_n est sans biais lorsque l'erreur commise en moyenne est nulle.

Propriété 1

Soit $(X_1, \ldots, X_n)$ un échantillon, et m l'espérance de la loi mère de l'échantillon[a]. Alors la moyenne empirique :

$$\overline{X_n} = \frac{1}{n} \sum_{k=1}^{n} X_k$$

est un estimateur sans biais de m.

[a] que l'on devrait noter plus rigoureusement $m(\theta)$ puisqu'il dépend du paramètre θ de la loi mère μ_θ.

Preuve.

□

Exercice. Les estimateurs $A_n = \dfrac{2}{n(n+1)} \displaystyle\sum_{k=1}^{n} kX_k$ et $B_n = 0$ de m sont-ils sans biais ?

Exercice. On suppose que la loi mère de l'échantillon est une loi uniforme $\mathscr{U}([0, \theta])$. On considère l'estimateur $T_n = \max(X_1, \ldots, X_n)$ de θ.

1. Montrer que T_n est une variable à densité et en déterminer une densité.

2. T_n est-il un estimateur sans biais de θ ? Donner un estimateur S_n sans biais de θ.

Simulation informatique. Simulons la situation précédente à l'aide de `Python` : on cherche à estimer le paramètre θ inconnu d'une loi uniforme $\mathscr{U}([0,\theta])$ à l'aide de l'estimateur S_n pour $n = 100$.

```
theta=rd.exponential(10) # choix de theta au hasard
X=theta*rd.random(100) # réalisation d'un échantillon de taille 100
S=(101/100)*np.max(X) # estimation de theta par S
print(S)
print(theta)
```

En exécutant ce programme, on obtient par exemple $S_{100} = 1,531$ alors que $\theta = 1,553$.

Notons qu'on n'est jamais à l'abri d'un « mauvais » échantillon : il se peut (même si ceci est peu probable, surtout lorsque n est grand) que tous les X_i soient par exemple inférieurs à $\dfrac{\theta}{2}$ et donc que S_n retourne une estimation de θ éloignée de sa valeur réelle.

3.2 Comparaison d'estimateurs sans biais

Afin de comparer différents estimateurs sans biais d'un même paramètre $g(\theta)$, et ainsi d'isoler les plus pertinents, c'est-à-dire ceux qui fourniront en moyenne les estimations les plus proches de $g(\theta)$, on comparera leur variance. En effet :

- si T_n est sans biais, mais que sa variance est grande, il prendra souvent des valeurs très éloignées de $g(\theta)$;
- à l'inverse, si T_n est sans biais et que sa variance est proche de 0, il donnera des estimations en moyenne très regroupées autour de $g(\theta)$.

Ainsi, le meilleur estimateur sans biais d'un paramètre $g(\theta)$ est celui qui à la variance la plus petite.

Méthode. Comparaison de deux estimateurs sans biais.

Pour comparer deux estimateurs sans biais T_n et U_n de $g(\theta)$, on calculera leur variance. Si on obtient :

$$\forall \theta \in \Theta, \quad V_\theta(T_n) < V_\theta(U_n),$$

alors T_n est un meilleur[a] *estimateur de $g(\theta)$ que U_n.*

[a] Il commet en moyenne moins d'erreurs pour estimer $g(\theta)$.

Exercice. Reprenons le cas d'une élection présidentielle, et donc d'un échantillon $(X_1, \ldots, X_n)$ de loi mère $\mathscr{B}(\theta)$. Des estimateurs $\overline{X_n}$ et $A_n = \dfrac{2}{n(n+1)} \displaystyle\sum_{k=1}^{n} kX_k$ de θ, lequel privilégieriez-vous ?

Simulation informatique. Testons nos résultats à l'aide de `Python`. Pour cela, on va choisir un paramètre θ au hasard dans $[0, 1]$, correspondant à la proportion réelle de votants pour le candidat A.

```
theta = rd.random()
```

`Python` connait déjà le résultat du scrutin, et on pourrait lui demander avec l'instruction `print(theta)`, mais gardons pour le moment le mystère... On va estimer ce paramètre à l'aide des résultats précédents. Il nous faut pour cela un échantillon observé, qu'on prendra de taille $n = 60$ par exemple.

```
e = rd.binomial(1,theta, 60)
```

Grâce à cet échantillon, on peut obtenir deux estimations $\overline{x_{60}}$ et a_{60} de θ à l'aide respectivement de $\overline{X_{60}}$ et A_{60}. Demandons à `Python` de les calculer :

```
x60 = np.mean(e) ; a60 = (2/(60*61))*np.sum(np.arange(1,61)*e)
print(x60,a60)
```

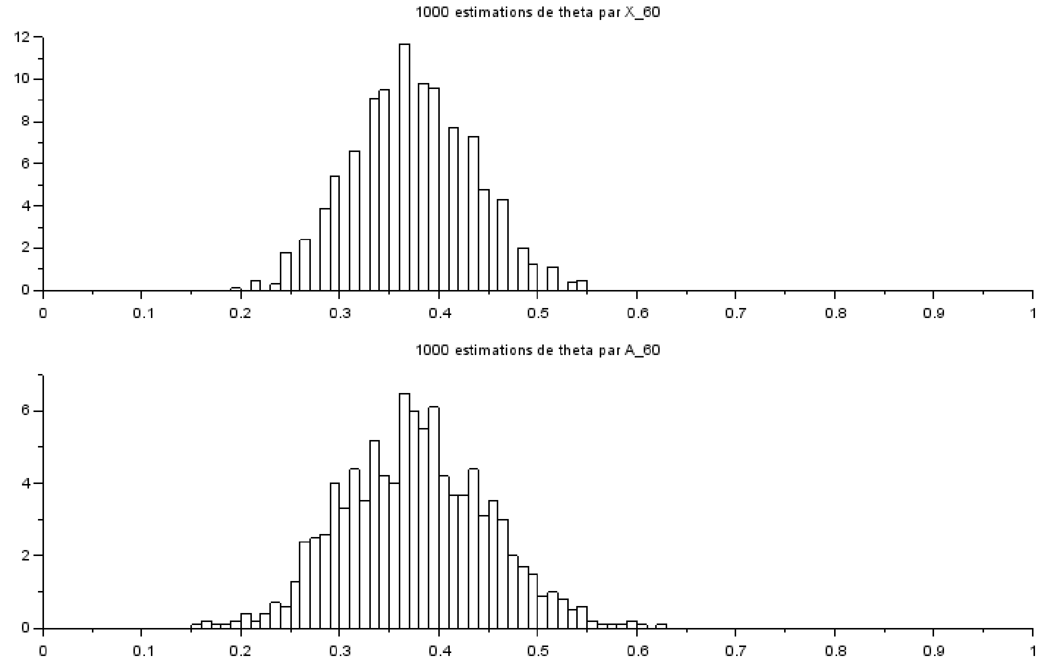
On obtient `x60 = 0.3166667` et `a60 = 0.2737705`. De ces deux estimations ponctuelles de θ, on privilégiera `x60` puisque, comme on l'a vu, $\overline{X_n}$ est un meilleur estimateur de θ que A_n. On peut enfin comparer ces estimations ponctuelles avec la valeur réelle de θ : `print(theta)` renvoie `0.3760119`. On remarque que `x60` est bien l'estimation la plus proche[6] de θ.

On peut répéter ce procédé pour 1000 échantillons de taille 60, et donc pour 1000 estimations de θ par $\overline{X_{60}}$ et A_{60}, et comparer les histogrammes des fréquences des estimations à l'aide du programme `Python` suivant :

```
n=60
E = rd.binomial(1,theta,[1000,n]) # 1000 échantillons observés
X=np.zeros(1000) # estimateur moyenne empirique
A=np.zeros(1000) # estimateur A_n
alpha=2/(n*(n+1))
for k in range(1000):
    X[k]=np.mean(E[k,:])
    A[k]=alpha*np.sum(np.arange(1,n+1)*E[k,:])

c=np.linspace(0,1,100)
plt.subplot(2,1,1)
plt.hist(X,c,density ='True',edgecolor='k')
plt.title('1000 estimations de theta par X_60')
plt.subplot(2,1,2)
plt.hist(A,c,density ='True',edgecolor='k')
plt.title('1000 estimations de m par A_60')
plt.show()
```

[6] Ce n'était pas nécessairement le cas, on aurait pu tomber ponctuellement sur un « mauvais » échantillon pour lequel la meilleure estimation aurait été `a60`, même si cela était peu probable.



Histogramme des fréquences des estimations de θ.

Notons que les estimations obtenues par $\overline{X_{60}}$ et par A_{60} sont en moyenne égales à θ (les histogrammes sont « centrés en θ »), ce qui n'est pas étonnant puisque ces deux estimateurs sont sans biais. On notera aussi que les estimations de θ obtenues à l'aide par A_{60} sont (très légèrement) plus dispersées que celles obtenues par $\overline{X_{60}}$, ce qui provient du fait que $V_\theta(\overline{X_{60}}) < V_\theta(A_{60})$ (ces deux termes étant malgré tout du même ordre de grandeur, ce n'est pas si probant que cela).

4 Suites d'estimateurs

4.1 Généralités

Définition.

Une *suite d'estimateurs* de $g(\theta)$ est la donnée pour tout $n \geq 1$ d'un estimateur T_n d'ordre n de $g(\theta)$. En particulier, chaque T_n est de la forme $\varphi_n(X_1, \dots, X_n)$.

Exemple. La suite $(\overline{X_n})_{n\geq 1}$ des moyennes empiriques est une suite d'estimateurs de l'espérance m de la loi mère.

Définition.

On dit que la suite $(T_n)_{n\in\mathbb{N}^*}$ d'estimateurs de $g(\theta)$ est *convergente* (ou plus simplement que T_n est un estimateur *convergent* de $g(\theta)$) lorsque pour tout $\theta \in \Theta$:

$$T_n \xrightarrow{P_\theta} g(\theta),$$

soit encore :

$$\forall \varepsilon > 0, \quad P_\theta(|T_n - g(\theta)| \geq \varepsilon) \underset{n\to+\infty}{\longrightarrow} 0.$$

Remarque. Cela signifie que pour n grand, T_n prend des valeurs très proches de $g(\theta)$ avec une probabilité voisine de 1.

Propriété 2

Si la loi mère admet une espérance m et une variance, alors $\overline{X_n} = \dfrac{1}{n}\displaystyle\sum_{k=1}^{n} X_k$ est un estimateur sans biais et convergent de m.

Preuve.

$\square$

Propriété 3

Si $(T_n)_{n\in\mathbb{N}^*}$ est une suite convergente d'estimateurs de $g(\theta)$ et si f est une fonction continue sur $\mathbb{R}$ à valeurs réelles, alors $\big(f(T_n)\big)_{n\in\mathbb{N}^*}$ est une suite convergente d'estimateurs de $f\big(g(\theta)\big)$.

4.2 Condition suffisante de convergence

Propriété 4 (Condition suffisante de convergence)

Soit $(T_n)_{n\geq 1}$ une suite d'estimateurs de $g(\theta)$ admettant des moments d'ordre 2 (ou des variances) pour tout $\theta \in \Theta$. **Si** pour tout $\theta \in \Theta$:

$$\lim_{n\to+\infty} E_\theta(T_n) = g(\theta) \quad \text{et} \quad \lim_{n\to+\infty} V_\theta(T_n) = 0,$$

alors T_n est un estimateur convergent de $g(\theta)$.

Preuve.

$\square$

Méthode. Comment montrer la convergence d'un estimateur ?

Pour montrer la convergence d'un estimateur T_n de $g(\theta)$, on pourra procéder comme suit :

- *regarder si cela ne relèverait pas directement de la loi faible des grands nombres ;*
- *sinon on peut calculer $E_\theta(T_n)$ et $V_\theta(T_n)$ et montrer qu'elles tendent respectivement vers $g(\theta)$ et 0 lorsque $n \to +\infty$;*
- *si toutes les méthodes précédentes échouent, on pourra tenter de calculer explicitement $P_\theta(|T_n - g(\theta)| \geq \varepsilon)$ pour tout $\varepsilon > 0$.*

Exercice. On suppose que la loi mère μ_θ admet une espérance m et une variance σ^2. L'estimateur $A_n = \dfrac{2}{n(n+1)} \displaystyle\sum_{k=1}^{n} kX_k$ est-il convergent ?

Exercice. On suppose que la loi mère de l'échantillon est $\mathscr{U}([0,\theta])$. On considère les estimateurs de θ suivants :

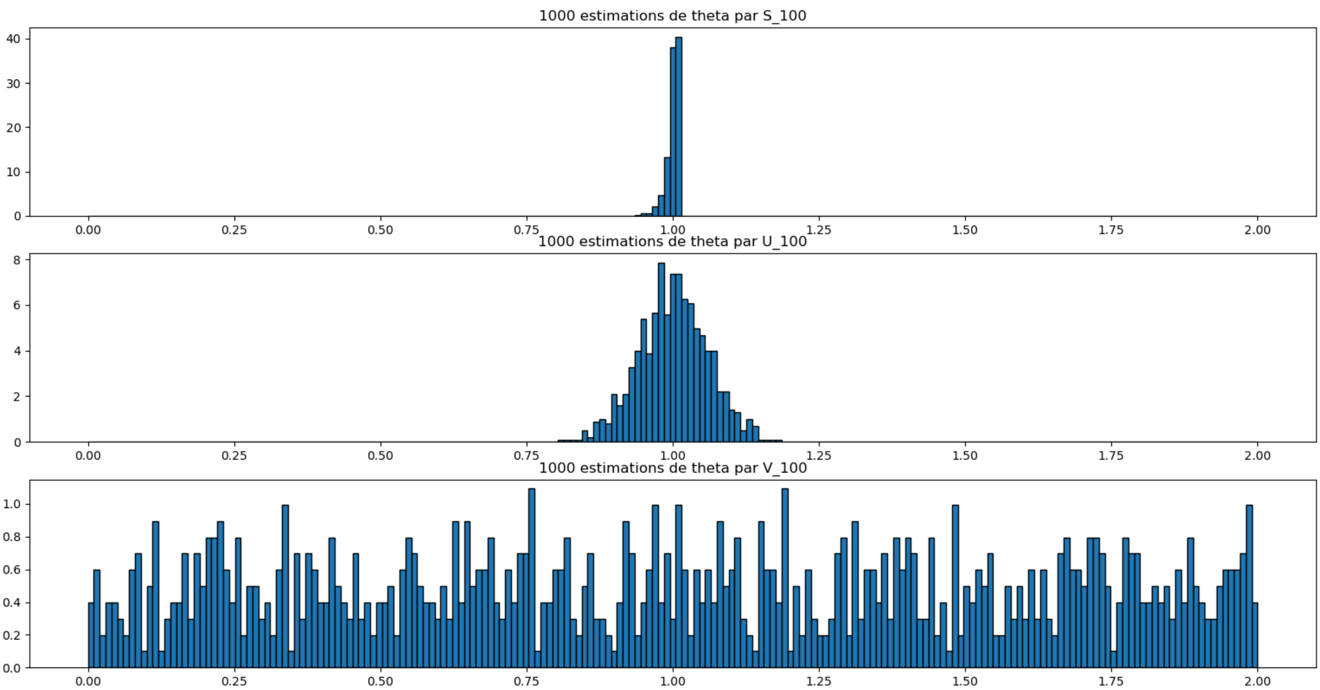
$$S_n = \frac{n+1}{n} \max(X_1, \dots, X_n) \quad , \quad U_n = 2\overline{X_n} \quad \text{et} \quad V_n = 2X_n.$$

1. Montrer que S_n est un estimateur convergent de θ.
2. Calculer l'espérance et la variance de U_n et de V_n. Ces estimateurs sont-ils sans biais ? Sont-ils convergents ?
3. Quel estimateur de θ privilégieriez-vous ?

Simulation informatique. On compare 1000 estimations de θ par S_{100}, U_{100} et V_{100} à l'aide du programme `Python` suivant (avec $\mathscr{U}([0,1])$ pour loi mère et donc $\theta_0 = 1$) :

```
n=100
E =rd.random([1000,n]) # 1000 échantillons observés
S=np.zeros(1000) # estimateur S_n
U=np.zeros(1000) # estimateur U_n
V=np.zeros(1000) # estimateur V_n
for k in range(1000):
    S[k]=((n+1)/n)*np.max(E[k,:])
    U[k]=2*np.mean(E[k,:])
    V[k]=2*E[k,n-1]

c = np.linspace(0,2,200)
plt.subplot(3,1,1)
plt.hist(S,c,density='True',edgecolor='k')
plt.title('1000 estimations de theta par S_100')
plt.subplot(3,1,2)
plt.hist(U,c,density='True',edgecolor='k')
plt.title('1000 estimations de theta par U_100')
plt.subplot(3,1,3)
plt.hist(V,c,density='True',edgecolor='k')
plt.title('1000 estimations de theta par V_100')
plt.show()
```



Histogramme des fréquences des estimations de θ.

Ces histogrammes confirment ce que nous avons établi par le calcul : l'estimateur S_n est celui qui fournit les meilleures estimations en moyenne parmi les trois estimateurs de θ considérés.