CHAPITRE 18

Estimations

If there is a 50-50 chance that something can go wrong, then nine times out of 10 it will.

PAUL HARVEY
Animateur radio américain (1918-2009)

1 Estimation ponctuelle

1.1 Principe : modéliser, estimer, tester/prédire

On considère un phénomène aléatoire et on s'intéresse à une variable aléatoire réelle X qui pourrait le décrire. On suppose que la loi de probabilité de X n'est pas complètement spécifiée et appartient à une famille de lois dépendant d'un paramètre θ décrivant un ensemble Θ. Le paramètre θ est une quantité inconnue, fixée dans toute l'étude, que l'on cherche à déterminer ou pour laquelle on cherche une information partielle.

Le problème de l'*estimation ponctuelle* consiste alors à préciser la vraie valeur du paramètre θ (ou plus généralement d'une fonction $g(\theta)$) à partir d'un échantillon de données $x_1, \dots, x_n$ obtenues en observant n fois le phénomène. Cette fonction du paramètre représentera en général une valeur caractéristique de la loi inconnue comme son espérance, sa variance, son étendue [1], etc.

Exemples.

- *Exemple 1. Nombre de buts en Ligue 1.*

→ Voici le nombre de buts par journée lors de la saison 2021/2022 :

26, 34, 29, 31, 28, 25, 32, 30, 25, 26, 29, 29, 29, 30, 21, 29, 29, 27,

27, 15, 26, 35, 30, 23, 23, 22, 21, 19, 23, 32, 38, 30, 26, 35, 30, 36, 30, 37.

Ces nombres constituent *l'échantillon de données*.

→ On pose un *modèle* en supposant que le nombre de buts lors d'une journée est une variable aléatoire X qui suit une loi de Poisson. Notons ce paramètre λ.

→ Pour donner une valeur à λ, (on dira *estimer*), on peut utiliser le fait que X admet une espérance $\mathbf{E}(X) = \lambda$. Ainsi, on peut préciser λ en prenant la valeur moyenne du nombre de buts (à savoir ici $\lambda = 2.81$).

→ Une fois le modèle posé, on peut le *tester* s'il donne des résultats cohérents et ainsi l'utiliser pour faire de la *prédiction*.

- *Exemple 2. Durée de vie d'un composant électrique.*

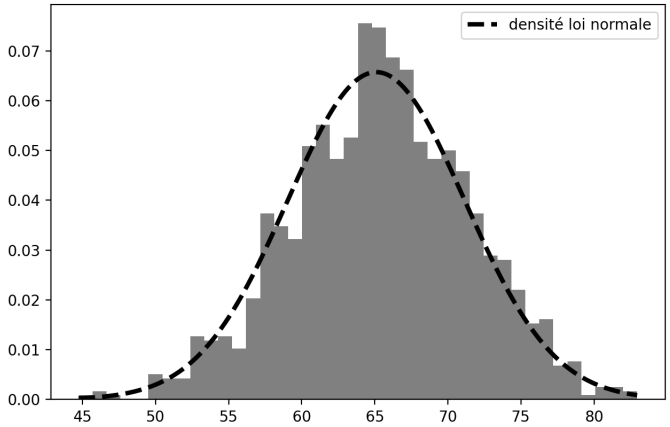
On suppose que l'emploi d'un composant électrique se fait dans des conditions normales d'utilisation et on néglige

1. Si $X(\Omega) = [a; b]$, l'étendue est définie par la quantité $b - a$.

les phénomènes d'usure. Autrement dit, on suppose que la variable aléatoire X qui donne la durée de vie du composant (en heure) est une loi sans mémoire. On a vu (voir exercice **??** p. **??**) que X suit alors une loi exponentielle. De plus, on appelle demi-vie de X le réel t tel que $\mathbf{P}(X \leqslant t) = \mathbf{P}(X \geqslant t)$ (on parle aussi de médiane). On vérifie que $t = \ln 2/\lambda$.
On teste une centaine de composants et au bout de 1000 heures d'utilisation, la moitié des composants ne fonctionnent plus. On fait alors le choix de λ tel que $1000 = \ln 2/\lambda$, soit $\lambda = \ln(2)/1000 \simeq 6.9 \cdot 10^{-4}$. À partir de ce modèle, on s'attend à ce que la durée de vie moyenne d'un composant est $\mathbf{E}(X) = 1/\lambda$, soit environ 1500 heures.

- *Exemple 3.*

Un producteur d'œufs de poules souhaite anticiper sa production pour l'année suivante en analysant sa production actuelle. Ci-contre, l'histogramme des différents poids (en grammes) des œufs récoltés dans l'année. Ce dernier modélise la situation à l'aide d'une loi normale. Les paramètres de la loi sont fixés à partir de la moyenne et de l'écart-type de l'échantillon. Dit autrement, si le poids d'un œuf correspond à une variable aléatoire X, on a $X \hookrightarrow \mathcal{N}(\mu;\sigma^2)$ avec $\mu = 65.1$ et $\sigma = 6.07$.



Console

```
# L représente la liste des
    différents poids

>>> len(L)
51243

>>> np.mean(L) # poids moyen
65.115928

>>> np.std(L)  # écart-type
6.0736541402592824
```

Partant de ce modèle, on peut calculer la probabilité d'un très gros œuf (poids supérieur à 73 grammes) :

$$\mathbf{P}(X > 73) = 1 - \mathbf{P}(X \leqslant 73) = 1 - \mathbf{P}\left(\frac{X-\mu}{\sigma} \leqslant \frac{73-65.1}{6.07}\right) \simeq 1 - \Phi(1.3) \simeq 10\%.$$

La suite de ce chapitre propose de donner un cadre théorique rigoureux à ces trois exemples et de juger aussi de la pertinence de cette approche. Nous discuterons essentiellement de l'étape d'estimation. Notamment nous verrons deux types d'estimations :

- ⇀ L'estimation ponctuelle .
- ⇀ L'estimation par intervalle de confiance.

1.2 Définitions et exemples

Dans la suite, on se fixe :

- un espace probabilisable $(\Omega, \mathscr{A})$.
- un espace des paramètres Θ qui est une partie de $\mathbb{R}^n$.
  On suppose de plus que pour chaque paramètre $\theta \in \Theta$, il existe une probabilité $\mathbf{P}_\theta$ définie sur $(\Omega, \mathscr{A})$.
- une application X qui est bien une variable aléatoire sur tous les espaces probabilisés $(\Omega, \mathscr{A}, \mathbf{P}_\theta)$ (où $\theta \in \Theta$).

Notation. Sous réserve d'existence :

- ⇀ $\mathbf{E}_\theta(X)$ désigne l'espérance de X pour la probabilité $\mathbf{P}_\theta$.
- ⇀ $\mathbf{V}_\theta(X)$ est la variance de X pour la probabilité $\mathbf{P}_\theta$.

Exemples.

- *Exemple 1.*

Dans ce cas, la variable X suit une loi de Poisson de paramètre θ. On a $\Theta = \mathbb{R}_*^+$ et

$$\forall \theta \in \mathbb{R}_*^+, \quad \forall n \in \mathbb{N}, \quad \mathbf{P}_\theta(X = n) = e^{-\theta}\frac{\theta^n}{n!}.$$

On a de plus, $\mathbf{E}_\theta(X_i) = \theta \quad$ et $\quad \mathbf{V}_\theta(X_i) = \theta$.

• *Exemple 2.* Dans le cas où X suit une loi exponentielle de paramètre $\lambda = \theta$, on a

$$\forall\, \theta \in \mathbb{R}_+^*, \quad \forall\, t \in \mathbb{R}_+^*, \quad \mathbf{P}_\theta(X \leqslant t) = 1 - e^{-\theta t}.$$

• *Exemple 3.* Dans ce dernier cas, X suit une loi normale $\mathcal{N}(\mu, \sigma^2)$ dont on ne connait ni l'espérance ni la variance, alors $\theta = (\mu, \sigma)$, et $\Theta = \mathbb{R} \times \mathbb{R}_+^*$.

DÉFINITION — **échantillon**

Soient $X : \Omega \to \mathbb{R}$ *une variable aléatoire définie sur* $(\Omega, \mathscr{A})$ *et* $n \in \mathbb{N}^*$.
On appelle **n-*échantillon*** *de la loi de* X *toute famille* $(X_1, \ldots, X_n)$ *telle que :*

- *Les applications* $X_1, \ldots, X_n$ *sont des variables aléatoires sur* $(\Omega, \mathscr{A})$.
- *Les variables* $X_1, \ldots, X_n$ *sont* $\mathbf{P}_\theta$*-indépendantes et de même loi que* X *pour tout* $\theta \in \Theta$.

Vocabulaire.
• On dit aussi que la loi de X est la loi parente (ou encore loi mère) de l'échantillon.
• On note souvent $(X_1, \ldots, X_n)$ *i.i.d* pour signaler que les variables sont indépendantes, et identiquement distribuées (c'est-à-dire de même loi).

En pratique, un échantillon de données $x_1, \ldots, x_n$ est la réalisation de n variables aléatoires $X_1, \ldots, X_n$. L'objectif de l'estimation ponctuelle est alors de déterminer le paramètre θ (ou une fonction $g(\theta)$) qui « explique » au mieux les valeurs de l'échantillon.

DÉFINITION — **estimateur**

• *On appelle* ***estimateur de*** θ *toute variable aléatoire de la forme* $\varphi(X_1, \ldots, X_n)$*, où* $(X_1, \ldots, X_n)$ *est un* n*-échantillon et* φ*, une fonction de* $\mathbb{R}^n$ *dans* $\mathbb{R}$.

• *Plus généralement, pour* $g : \Theta \to \mathbb{R}$*, une fonction, un* ***estimateur de*** $\mathbf{g}(\theta)$ *est une variable aléatoire de la forme* $\varphi(X_1, \ldots, X_n)$ *où* $(X_1, \ldots, X_n)$ *est un* n*-échantillon.*

Remarques.
• Un estimateur ne peut pas dépendre de θ puisque c'est la valeur que l'on souhaite déterminer.
• Estimer ponctuellement $g(\theta)$ par $\varphi(x_1, \ldots, x_n)$ où $\varphi(X_1, X_2, \ldots, X_n)$ est un estimateur de $g(\theta)$ et $(x_1, \ldots, x_n)$ est une réalisation de l'échantillon $(X_1, \ldots, X_n)$, c'est décider d'accorder à $g(\theta)$ la valeur $\varphi(x_1, \ldots, x_n)$.

Exemples.
• Avec les notations de la définition, soit $(X_1, \ldots, X_n)$ un n-échantillon de la loi de X. Il est souvent utile de considérer **l'estimateur de la moyenne empirique** donné par :

$$\overline{X_n} = \frac{X_1 + X_2 + \cdots + X_n}{n} \qquad (\bullet)$$

Mais on peut inventer toute une gamme d'estimateurs :

$$T_n = \max_{1 \leqslant k \leqslant n} X_k, \quad U_n = \min_{1 \leqslant k \leqslant n} X_k, \quad V_n = \ln(1 + |U_n|), \quad W_n = \max\{k \in [\![1, n]\!] \mid X_k = T_n\}, \quad Y_n = 1, \quad \text{etc.}$$

Editeur

```
theta=np.random.rand()
theta=0.4292081919392057 # le paramètre "inconnu"

n=500 # taille de l'échantillon
Ech=theta*np.random.rand(n)
# création de l'échantillon

np.round(Ech,2)
array([0.18, 0.25, 0.28, 0.26, 0.05, 0.14, 0.15, 0.02,
    0.15, 0.35, 0.2 ,
       0.25, 0.41, 0.17, 0.39, 0.32, 0.22, 0.21, 0.12,
           0.14, 0.35, 0.18,
       0.16, 0.23, 0.1 , 0.19, 0.28, 0.3 , 0.31, 0.23, 0.4
            , 0.08, 0.18 ...
```

• Créons un échantillon d'une loi uniforme $[0;\theta]$.

On peut essayer de retrouver la valeur de θ à partir de deux estimateurs

$$2\overline{X_n} \quad \text{et} \quad T_n = \max_{1\leqslant k\leqslant n} X_k.$$

On obtient deux estimations de θ par :

```
>>> 2*sum(Ech)/500
0.4386345472617729

>>> max(Ech)
0.4289769082767542
```

1.3 Biais, convergence et comparaison des estimateurs

Les définitions qui suivent permettent de quantifier la « qualité » d'un estimateur et de les comparer entre-eux.

DÉFINITION — **biais d'un estimateur**

Soit T_n un estimateur de $g(\theta)$ tel que tout $\theta\in\Theta$, T_n admet une espérance pour la probabilité $\mathbf{P}_\theta$.

- *On définit le **biais** de T_n en $g(\theta)$ par*
$$b_\theta(T_n) = \mathbf{E}_\theta(T_n) - g(\theta).$$
- *Si pour tout $\theta\in\Theta$, $b_\theta(T_n)=0$, on dit l'estimateur est **sans biais**. Sinon, on dit que l'estimateur est biaisé.*

Exemples.

• Avec les notations de la définition et si X admet une espérance θ, l'estimateur de la moyenne empirique $\overline{X_n}$ est un estimateur sans biais de θ.
En effet, pour tout $i\in[\![1,n]\!]$, on a $\mathbf{E}_\theta(X_i)=\theta$. Par linéarité de l'espérance $\mathbf{E}_\theta$, il vient

$$\mathbf{E}_\theta\left(\overline{X_n}\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{E}_\theta(X_i) = \theta, \quad \text{puis} \quad b_\theta\left(\overline{X_n}\right) = \mathbf{E}_\theta\left(\overline{X_n}\right) - \theta = 0.$$

• Considérons une variable X suivant une loi uniforme sur $[0,\theta]$. Pour tout $n\in\mathbb{N}^*$, posons $M_n=\max(X_1,\dots,X_n)$. On vérifie que M_n est une variable à densité, et une densité est donnée par

$$\forall\, x\in\mathbb{R}, \quad f_\theta(x) = \begin{cases} n\dfrac{x^{n-1}}{\theta^n} & \text{si } x\in[0;\theta] \\ 0 & \text{sinon.}\end{cases}$$

En tant que variable bornée, M_n admet une espérance avec

$$\mathbf{E}_\theta(M_n) = \int_0^\theta x f_\theta(x)\,\mathrm{d}x = \int_0^\theta n\frac{x^n}{\theta^n}\,\mathrm{d}x = \frac{n}{n+1}\theta.$$

On en déduit que M_n est un estimateur biaisé de θ et le biais de M_n est

$$b_\theta(M_n) = \frac{n}{n+1}\theta - \theta = -\frac{\theta}{n+1}.$$

Par contre, par linéarité de l'espérance, $\widetilde{M_n} = \frac{n+1}{n}M_n$ est un estimateur sans biais de θ puisque

$$\mathbf{E}_\theta\left(\widetilde{M_n}\right) = \frac{n+1}{n}\mathbf{E}_\theta(M_n) = \theta.$$

Remarque. On peut donner une définition moins contraignante : un estimateur est **asymptotiquement sans biais** si

$$b_\theta(T_n) \underset{n\to\infty}{\longrightarrow} 0.$$

Par exemple, la variable M_n définie précédemment est un estimateur asymptotiquement sans biais.

Exercice 1

✦✦ Soit $(X_1,\dots,X_n)$ un n-échantillon de la loi de Bernoulli de paramètre p. On pose

$$S_n = \sum_{i=1}^{n} X_i \quad \text{et} \quad T_n = \frac{S_n}{n}\left(1-\frac{S_n}{n}\right).$$

p. ??

1. Déterminer $\mathbf{E}(S_n)$ et $\mathbf{E}\left({S_n}^2\right)$. En déduire $\mathbf{E}(T_n)$.
2. À l'aide de T_n, proposer un estimateur sans biais de la variance de cette loi de Bernoulli.

Exercice 2

✦✦ ✎ **Estimateurs de la variance**

Soit la variable X admet un espérance μ et une variance σ^2. On pose

$$T_n = \frac{1}{n}\sum_{k=1}^{n}(X_k-\mu)^2 \quad \text{et} \quad V_n = \frac{1}{n}\sum_{k=1}^{n}\left(X_k-\overline{X_n}\right)^2.$$

p. 23

1. On suppose, dans cette question, que μ est connu et on cherche à estimer σ^2 qui est donc inconnue. Montrer que T_n est un estimateur sans biais de σ^2.
2. On suppose maintenant que μ est aussi inconnu.
   a) Montrer que V_n est un estimateur asymptotiquement sans biais de σ^2 et calculer le biais de cet estimateur.
   b) Construire, à partir de V_n, un estimateur sans biais de σ^2.

DÉFINITION — estimateur convergent

Soit $(T_n)_{n\in\mathbb{N}^*}$, une suite d'estimateurs de $g(\theta)$.
*On dit que la suite $(T_n)_{n\in\mathbb{N}}$ est **convergente** si pour tout $\theta\in\Theta$, la suite (T_n) converge en probabilité vers la variable aléatoire presque sûrement constante $g(\theta)$. Autrement dit*

$$\forall\,\theta\in\Theta,\quad \forall\varepsilon\in\mathbb{R}_*^+,\qquad \mathbf{P}_\theta\left(\left|T_n-g(\theta)\right|\geqslant\varepsilon\right)\underset{n\to\infty}{\longrightarrow}0.$$

Vocabulaire. Par abus de langage, on dit simplement que T_n est un estimateur convergent de $g(\theta)$.

Exercice 3

✦ **Exemples**

1. Considérons X une variable aléatoire dont la loi est uniforme sur $[0,\theta]$, où le paramètre θ est inconnu. Vérifier que $T_n = 2\overline{X}_n$ est un estimateur convergent de θ.
2. Soit X une variable à densité donnée pour $\theta\in\mathbb{R}_*^+$ par

$$\forall\,x\in\mathbb{R},\quad f_\theta(x)=\frac{2}{\theta}\left(1-\frac{x}{\theta}\right)\mathbf{1}_{[0,\theta]}(x).$$

p. 23

   a) Pour un échantillon $(X_1,\dots,X_n)$, montrer que $3\overline{X_n}$ est un estimateur sans biais de θ.
   b) Vérifier que cet estimateur est convergent.

Remarque. La notion de convergence des estimateurs ne donne aucune assurance pratique que la valeur prise par un estimateur à partir de l'échantillon de données sera assez proche de la vraie valeur du paramètre. On quantifie la qualité des estimateurs par la notion de **risque quadratique**. Cette notion est maintenant hors-programme mais on pourra consulter l'exercice 12, page 15, pour des précisions.

PROPOSITION — composition et estimateurs

Soit $(T_n)_{n\in\mathbb{N}}$, *une suite d'estimateurs de* $g(\theta)$.

Si | → *La suite* $(T_n)_{n\in\mathbb{N}}$ *est une suite d'estimateurs convergente de* $g(\theta)$.
| → *La fonction* f *est continue sur* $\mathbb{R}$.

Alors $\big(f(T_n)\big)_{n\in\mathbb{N}}$ *est une suite d'estimateurs convergente de* $f\big(g(\theta)\big)$.

Preuve. C'est une conséquence directe de l'énoncé : Soit $(X_n)_{n\in\mathbb{N}}$ une suite de variables aléatoires sur $(\Omega,\mathscr{A},\mathbf{P})$.

Si | → La suite $(X_n)_{n\in\mathbb{N}}$ converge en probabilité vers X.
| → La fonction f est continue sur $\mathbb{R}$ à valeurs réelles.

Alors $f(X_n) \xrightarrow[n\to+\infty]{\mathbf{P}} f(X)$.

■

Exemples. Nous en donnons deux.

• Reprenons le cas de X une variable aléatoire dont la loi est uniforme sur $[0;\theta]$ avec θ inconnu. À partir du théorème de transfert, on vérifie que $\ln(X)$ a une espérance avec

$$\mathbf{E}_\theta(\ln(X)) = \int_0^\theta \ln(t)\,\mathrm{d}t = \ln(\theta) - 1.$$

On vérifie de même que $\ln(X)$ admet un moment d'ordre 2 (et donc une variance). Toujours d'après la loi faible des grands nombres, $\left(\sum_{i=1}^{n} \ln(X_i)\right)/n$ est un estimateur convergent de $\ln(\theta)-1$. La fonction $f : t\in\mathbb{R}\mapsto \exp(t+1)$ est continue, donc l'estimateur suivant est un estimateur convergent de θ :

$$\mathrm{e}\cdot\sqrt[n]{\prod_{i=1}^{n} X_i} = \exp\left(\frac{\sum_{i=1}^{n}\ln(X_i)}{n}+1\right) = f\left(\frac{1}{n}\sum_{i=1}^{n}\ln(X_i)\right) \xrightarrow[n\to+\infty]{\mathbf{P}} f\big(\ln(\theta)-1\big) = \theta.$$

• Soit $(X_n)_{n\in\mathbb{N}^*}$ des variables aléatoires indépendantes de même loi géométrique $\mathscr{G}(p)$. Comme la loi géométrique est d'espérance $1/p$ (et admettant une variance), la loi faible justifie que $\sum_{i=1}^{n} X_i/n$ est un estimateur de $1/p$. Comme la fonction inverse est continue sur $\mathbb{R}_*^+$, $n/\left(\sum_{i=1}^{n} X_i\right)$ est un estimateur convergent de p.

Notons que la conclusion du théorème demeure si f est continue sur un intervalle I contenant $X(\Omega)$.

Exercice 4

✦ Soit $(X_1,X_2,\dots,X_n)$ un échantillon de la loi normale $\mathscr{N}(0,\sigma^2)$.

1. Pour tout $i\in[\![1;n]\!]$, calculer l'espérance de la variable aléatoire $|X_i|$.

2. En déduire un estimateur sans biais et convergent de σ.

p. 24

Exercice 5

✦✦ **Biais et composition**

Soit $(X_1,X_2,\dots,X_n)$ un échantillon de la loi de Poisson de paramètre λ inconnu. On sait que la moyenne empirique $\overline{X}_n = \sum_{k=1}^{n} X_k/n$ est un estimateur sans biais et convergent de λ. On cherche à estimer $\mathrm{e}^{-\lambda}$.

Est-ce que l'estimateur $T_n = \mathrm{e}^{-\overline{X}_n}$ est un estimateur sans biais de $\mathrm{e}^{-\lambda}$? asymptotiquement sans biais ? convergent ?

p. 24

PROPOSITION — condition suffisante de convergence

Soit $(T_n)_{n\in\mathbb{N}}$, *une suite d'estimateurs de* $g(\theta)$.

Si $\quad \mathbf{E}(T_n) \underset{n\to\infty}{\longrightarrow} g(\theta) \quad$ *et* $\quad \mathbf{V}(T_n) \underset{n\to\infty}{\longrightarrow} 0.$

Alors $(T_n)_{n\in\mathbb{N}}$ *est une suite d'estimateurs convergente de* $g(\theta)$.

Exercice 6

✧ Prouver cet énoncé. p. 24

Remarque. Un estimateur sans biais Y_n est meilleur qu'un autre estimateur sans biais Z_n si $\mathbf{V}(Y_n) \leqslant \mathbf{V}(Z_n)$ pour tout entier n. Le fait d'être meilleur se matérialise dans l'inégalité de Bienaymé-Tchebychev :

$$\forall \varepsilon \in \mathbb{R}_*^+, \quad \mathbf{P}_\theta\left(\left|T_n - g(\theta)\right| \geqslant \varepsilon\right) \leqslant \frac{\mathbf{V}(Y_n)}{\varepsilon^2} \leqslant \frac{\mathbf{V}(Z_n)}{\varepsilon^2}.$$

Exemple. L'estimateur de la moyenne empirique est convergent si X admet un variance.

Exercice 7

✧ Soit $(X_n)_{n\geqslant 1}$ une suite de variables aléatoires mutuellement indépendantes et suivant toutes la loi $\mathscr{B}(p)$, où p est un paramètre inconnu que l'on cherche à estimer. On pose

$$\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{et} \quad T_n = \frac{2}{n(n+1)}\sum_{i=1}^{n} i X_i.$$

p. 25

1. Montrer que $\overline{X_n}$ et T_n sont deux estimateurs sans biais de p.
2. Calculer et comparer les variances de $\overline{X_n}$ et de T_n.
3. Montrer que $\overline{X_n}$ et T_n sont deux estimateurs convergents de p.

Exemple. Questions sensibles lors d'un sondage d'opinion.
Certains sujets abordés dans les sondages d'opinion peuvent être sensibles et les personnes interrogées peuvent refuser de répondre honnêtement. Détaillons une procédure qui permet aux sondés de répondre plus librement. Considérons n personnes sondées et une question fermée à deux réponses possibles dont on veut estimer la probabilité p de réponses positives dans la population générale. On demande à chaque sondé de lancer un dé.

- S'il obtient 6, la personne doit donner sa réponse sans mentir.
- Sinon, elle donne la réponse contraire à la sienne.

Si le sondeur ignore le résultat du dé, il ne pourra pas savoir si la réponse est franche ou non, et on peut espérer que la personne sondée acceptera plus facilement de répondre honnêtement à la question.
Généralisons la procédure en fixant t, la probabilité que la personne réponde sans mentir. Le réel t est connu, il vaut $1/6$ dans l'exemple du dé. Posons pour tout $i \in [\![1;n]\!]$, la variable aléatoire X_i valant 1 le i-ème sondé répond positivement et 0 sinon. X_i suit une loi de Bernoulli. Donnons son paramètre à l'aide de la formule de probabilités totales avec le système complet d'événements constitué de A :« Le i sondé ne ment pas » et $\overline{A}$.

$$\mathbf{P}(X_i = 1) = \mathbf{P}(A)\mathbf{P}_A(X_i = 1) + \mathbf{P}(\overline{A})\mathbf{P}_{\overline{A}}(X_i = 1) = tp + (1-t)(1-p).$$

Nous avons vu que l'estimateur de la moyenne empirique $\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$ est un estimateur sans biais et convergent de $r := tp + (1-t)(1-p)$. Or, pour $t \neq 1/2$, on peut inverser la relation

$$r = tp + (1-t)(1-p) \iff p = \frac{1-r-t}{2t-1} \iff p = f(r) \quad \text{où} \quad f : x \in \mathbb{R} \mapsto \frac{1-x-t}{2t-1}.$$

Posons alors :

$$T_n = f(\overline{X_n}) = \frac{1-t-\overline{X_n}}{1-2t}.$$

T_n est alors est estimateur convergent de p. Donnons deux arguments :

→ La fonction f est continue et $\overline{X_n}$ est un estimateur convergent de r. Donc $T_n = f(\overline{X_n})$ est un estimateur convergent de $f(r) = p$.

→ Par linéarité de l'espérance, on vérifie que l'espérance de T_n est p . T_n est un estimateur sans biais de p. De plus, la variance de T_n vaut :

$$\mathbf{V}(T_n) = \frac{r(1-r)}{n(2t-1)^2} \underset{n\to\infty}{\longrightarrow} 0.$$

D'après la proposition précédente, on retrouve le fait que l'estimateur T_n est convergent de p.

Comparaison des estimateurs sans biais

Simulation Python.
Reprenons le cas d'une variable X suivant une loi uniforme sur $[0;\theta]$ et des deux estimateurs sans biais :

$$\overline{X_n} = \frac{2}{n}\sum_{i=1}^{n} X_i \quad \text{et} \quad \widetilde{M_n} = \frac{n+1}{n}\max(X_1,\ldots,X_n).$$

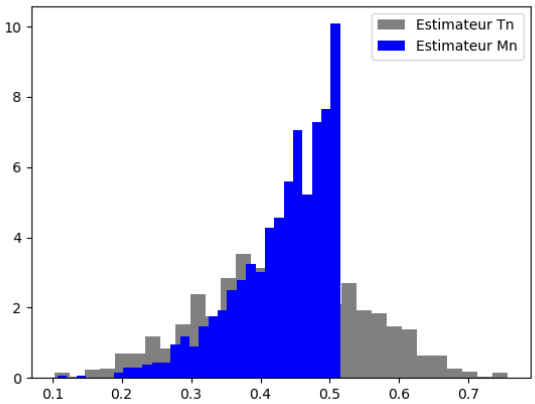
Simulons un grand nombre d'échantillons de données et affichons les réalisations de ces deux estimateurs à l'aide des histogrammes.

Editeur

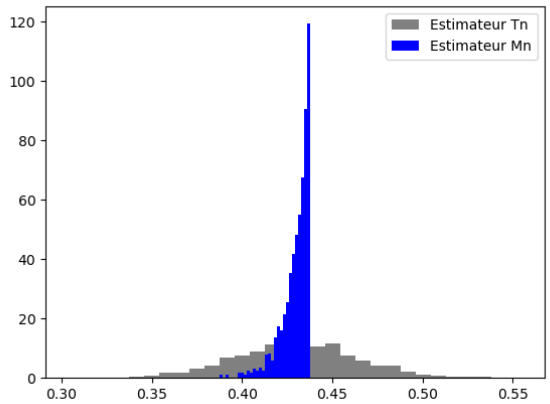
```
theta=0.4292081919392057
# choix du paramètre "inconnu"

def estimateurT(n):
    E=theta*np.random.rand(n)
    return 2*np.sum(E)/n

def estimateurM(n):
    return ((n+1)/n)*max(theta*np.random
        .rand(n))
```

```
# affichage des histogrammes

def histogrammes(n):
    LM=np.zeros(1000)
    LT=np.zeros(1000)
    for i in range(1000):
        LM[i]=estimateurM(n)
        LT[i]=estimateurT(n)

    plt.hist(LM,30)
    plt.hist(LT,30)
    plt.show()
```

```
>>> histogrammes(5)
```




```
>>> histogrammes(50)
```

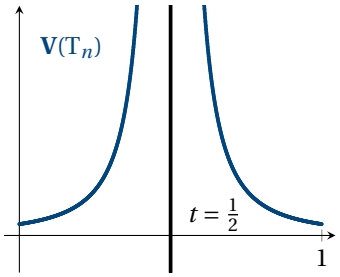


Le meilleur estimateur sans biais est celui qui a la variance la plus faible. En effet, c'est celui où le risque que l'échantillon de données donne une valeur loin de l'espérance (qui est dans ce cas le paramètre à estimer) est le plus faible.

Exemple. On a vu que dans le cas de questions sensibles lors d'un sondage d'opinion, la variance de l'estimateur est

$$\mathbf{V}(\mathrm{T}_n) = \frac{r(1-r)}{n(2t-1)^2}.$$

On retrouve bien le cas que l'estimateur est meilleur lorsque $t = 0$ (le sondé ne ment jamais) ou $t = 1$ (le sondé ment systématiquement).



2 Estimation par intervalle de confiance

S'il existe des critères pour juger des qualités d'un estimateur ponctuel T_n de $g(\theta)$, aucune certitude ne peut jamais être apportée quant au fait que l'estimation donnée par l'échantillon de données soit une « bonne » valeur du paramètre $g(\theta)$. L'estimation par intervalle de confiance permet de trouver un intervalle aléatoire qui contienne $g(\theta)$ avec une probabilité minimale donnée. Dans tout ce paragraphe, $(\mathrm{U}_n)_{n\in\mathbb{N}^*}$ et $(\mathrm{V}_n)_{n\in\mathbb{N}^*}$ désigneront deux suites d'estimateurs de $g(\theta)$ telles que pour tous $\theta \in \Theta$ et $n \in \mathbb{N}^*$,

$$\mathbf{P}_\theta\left([\mathrm{U}_n \leqslant \mathrm{V}_n]\right) = 1.$$

2.1 Intervalle de confiance

Définition

DÉFINITION — **intervalle de confiance**

Soient $\alpha \in]0;1[$, U_n *et* V_n *deux estimateurs de* $g(\theta)$ *tels que pour tout* $\theta \in \Theta$

$$\mathbf{P}_\theta\Big(\, g(\theta) \in \left[\mathrm{U}_n;\mathrm{V}_n\right] \Big) \geqslant 1-\alpha.$$

On dit que l'intervalle $[\mathrm{U}_n;\mathrm{V}_n]$ *est un* ***intervalle de confiance*** *de* $g(\theta)$ *avec un risque d'au plus* α *ou au niveau de confiance au moins égal à* $1-\alpha$.

Remarques.

- En pratique, on part d'un échantillon de données $x_1, x_2, \ldots, x_n$. On calcule les valeurs

$$u_n = \mathrm{U}_n(x_1, x_2, \ldots, x_n) \quad \text{et} \quad v_n = \mathrm{V}_n(x_1, x_2, \ldots, x_n).$$

Ainsi on construit un intervalle aléatoire $[u_n; v_n]$ dans lequel $g(\theta)$ à une probabilité supérieure à $1-\alpha$ de s'y trouver.

- Dans les conditions usuelles, on considère des niveaux de confiance de 95% ou 99% (soit $\alpha = 0,05$ ou $\alpha = 0,01$).

Estimation par intervalle de confiance en utilisant l'inégalité de Bienaymé-Tchebychev

Méthode

Comment obtenir un intervalle de confiance à partir de l'inégalité de Bienaymé-Tchebychev?

Soit T_n, une variable aléatoire admettant une variance, l'inégalité s'écrit

$$\mathbf{P}_\theta\Big(|T_n - \mathbf{E}_\theta(T_n)| \geqslant \varepsilon\Big) \leqslant \frac{\mathbf{V}_\theta(T_n)}{\varepsilon^2}.$$

Comme $\Big[|T_n - \mathbf{E}_\theta(T_n)| < \varepsilon\Big] \subset \Big[[|T_n - \mathbf{E}_\theta(T_n)| \leqslant \varepsilon]\Big]$, on obtient par passage au complémentaire

$$\mathbf{P}_\theta\Big(|T_n - \mathbf{E}_\theta(T_n)| \leqslant \varepsilon\Big) \geqslant 1 - \frac{\mathbf{V}_\theta(T_n)}{\varepsilon^2}.$$

Si T_n est un estimateur sans biais de $g(\theta)$ (c'est-à-dire $\mathbf{E}_\theta(T_n) = g(\theta)$) et si on peut trouver un entier n tel que $\frac{\mathbf{V}_\theta(T_n)}{\varepsilon^2} \leqslant \alpha$, on peut récrire l'inégalité

$$\mathbf{P}_\theta\left(T_n - \varepsilon \leqslant g(\theta) \leqslant T_n + \varepsilon\right) \geqslant 1 - \alpha.$$

L'intervalle de confiance est alors $[T_n - \varepsilon; T_n + \varepsilon]$, il est de longueur 2ε.

Exemple. *Estimation du paramètre p d'une loi de Bernoulli.*
On réalise un sondage sur n personnes avec une unique question. On suppose que les réponses des personnes sont indépendantes et on veut déterminer un intervalle de confiance d'au moins 0.95 de la probabilité p de répondre positivement à l'unique question posée. D'après la loi faible des grands nombres, cet intervalle sera d'autant plus petit (au sens de l'inclusion) que le nombre de personnes interrogés sera grand. Désignons, pour un entier naturel non nul quelconque n, par X_n la variable aléatoire égale à 1 si la $n^{\text{ième}}$ personne répond positivement et 0 sinon. La variable X_n suit une loi de Bernoulli, de paramètre p. En particulier

$$\mathbf{E}(X_n) = p \quad \text{et} \quad \mathbf{V}(X_n) = p(1-p).$$

Soit $\overline{X}_n$ la moyenne empirique

$$\overline{X}_n = \frac{1}{n}\sum_{k=1}^{n} X_k, \quad \mathbf{E}\left(\overline{X}_n\right) = p \quad \text{et} \quad \mathbf{V}(X_n) = \frac{p(1-p)}{n} \quad \text{(indépendances)}.$$

D'après l'inégalité de Bienaymé-Tchebychev :

$$\mathbf{P}\left(\left|\overline{X}_n - p\right| \leqslant \varepsilon\right) \geqslant 1 - \frac{p(1-p)}{n\varepsilon^2}.$$

Par une étude de fonction, on peut majorer $p(1-p)$ par $\frac{1}{4}$. On obtient

$$\mathbf{P}\left(\left|\overline{X}_n - p\right| \leqslant \varepsilon\right) \geqslant 1 - \alpha \quad \text{et} \quad \alpha = \frac{1}{4n\varepsilon^2} \iff \varepsilon = \frac{1}{2\sqrt{n\alpha}}.$$

Dans ce cas, on obtient :

$$\mathbf{P}\left(p \in \left[\overline{X}_n - \frac{1}{2\sqrt{n\alpha}}, \overline{X}_n + \frac{1}{2\sqrt{n\alpha}}\right]\right) \geqslant 1 - \alpha$$

On obtient ainsi un intervalle de confiance de p à un niveau de confiance $1-\alpha$ avec

$$\left[\overline{X}_n - \frac{1}{2\sqrt{n\alpha}}, \overline{X}_n + \frac{1}{2\sqrt{n\alpha}}\right].$$

Pour $\alpha = 0.05$, on a en particulier

$$\mathbf{P}\left(p \in \left[\overline{X}_n - \frac{1}{\sqrt{0.05n}}, \overline{X}_n + \frac{1}{\sqrt{0.05n}}\right]\right) \geqslant 0.95$$

et si on prend maintenant $n = 100$, on obtient $\varepsilon \simeq 0.22$. Donc $\left]\overline{X_n} - 0.22, \overline{X_n} + 0.22\right[$ est un intervalle de confiance de p au niveau de risque 0.05. Autrement dit, il y a plus de 95% de chances que p soit compris entre $\overline{X}_n - 0.22$ et $\overline{X_n} + 0.22$. L'étendue de l'intervalle est énorme (0.44) lorsque qu'on se rappelle que $p \in [0;1]$.
À l'inverse, on peut fixer ε, et s'intéresser à la taille de l'échantillon nécessaire pour garantir que l'étendue de l'intervalle $\left]\overline{X_n} - \varepsilon, \overline{X_n} + \varepsilon\right[$ soit inférieure à 1%, il faut

$$n \geqslant \frac{1}{4 \times 0.05 \times \varepsilon^2} = 50000.$$

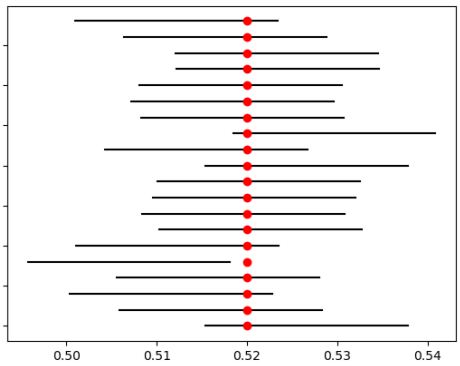
On constate qu'il faut interroger 50 000 personnes!

Remarques.
- Notons aussi que pour diviser par 2 la longueur de l'intervalle, il faut quadrupler le nombre de sondés.
- On retrouve le fait que l'inégalité de Bienaymé-Tchebychev est trop générale pour donner des résultats précis : elle ne prend pas suffisamment en compte la loi de $\overline{X_n}$, seules son espérance et sa variance sont importantes. Nous verrons dans la suite comment construire des intervalles de confiance plus petits en utilisant plus finement la loi de $\overline{X_n}$.

Simulation Python. Le code suivant crée plusieurs séries de données et construit l'intervalle aléatoire associé à chaque série de données.

Editeur

```
p=0.52
# le paramètre "inconnu" à estimer

n=10000 # Nombre de sondés
alpha=0.20 # niveau de risque

NbreTest=20
for i in range(NbreTest):
    # Création de l'échantillon de données
    Ech=np.random.rand(n)<p
    # Calcul des bords de l'intervalle
    u=np.mean(Ech)-1/(2*(n*alpha)**(1/2))
    v=np.mean(Ech)+1/(2*(n*alpha)**(1/2))
    plt.plot([u,v],[i,i])
    plt.plot([p],[i],'o')
plt.show()
```



On constate que dans la majorité des cas, le paramètre inconnu est bien dans l'intervalle aléatoire construit à partir de l'échantillon.

Estimation par intervalle de confiance de la moyenne d'une loi normale dont l'écart type est connu

Soit $(X_1, X_2, \ldots, X_n)$ un n-échantillon issu d'une loi normale $\mathcal{N}(\mu, \sigma^2)$. On suppose σ connu mais l'espérance μ est inconnue et on cherche à l'estimer. On considère $\overline{X_n}$ la moyenne empirique de l'échantillon. Nous avons vu que c'est un estimateur sans biais et convergent de μ. De plus, par les règles de stabilités des lois normales indépendantes :

$$\overline{X}_n \hookrightarrow \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

On a aussi $$Y_n = \sqrt{n} \cdot \frac{\overline{X}_n - \mu}{\sigma} \hookrightarrow \mathcal{N}(0,1).$$

Posons t_α positif tel que $t_\alpha = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$. Avec ce choix

$$\mathbf{P}(-t_\alpha \leqslant Y_n \leqslant t_\alpha) = \Phi(t_\alpha) - \Phi(-t_\alpha) = 2\Phi(t_\alpha) - 1 = 1 - \alpha.$$

En revenant à $\overline{X}_n$, on trouve

$$\mathbf{P}\left(\overline{X}_n - t_\alpha \frac{\sigma}{\sqrt{n}} \leqslant \mu \leqslant \overline{X}_n + t_\alpha \frac{\sigma}{\sqrt{n}}\right) = \mathbf{P}\left(-t_\alpha \leqslant \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \leqslant t_\alpha\right) = \mathbf{P}(-t_\alpha \leqslant Y_n \leqslant t_\alpha) = 1 - \alpha$$

On vient de montrer que

$$\left[\overline{X}_n - t_\alpha \frac{\sigma}{\sqrt{n}}, \overline{X}_n + t_\alpha \frac{\sigma}{\sqrt{n}}\right]$$

est un intervalle de confiance de μ avec un niveau de confiance égal à 0,95.

Applications numériques

- Pour un risque $\alpha = 5\%$, $1 - \frac{\alpha}{2} = 0,975$ et $t_{0.05} = \Phi^{-1}(0,975) \simeq 1,96$.
- Pour un risque de $\alpha = 1\%$, $1 - \frac{\alpha}{2} = 0,995$ et $t_{0.01} = \Phi^{-1}(0,995) \simeq 2,58$.

Exercice 8

✦✦ **Comparaison des méthodes précédentes**

Dans une population donnée, une étude statistique faite sur un groupe de 100 personnes donne lieu à la série statistique suivante.

Poids	48	49	50	51	52	53	54	55
Effectif	3	5	2	6	6	10	12	10
Poids	56	57	58	59	60	61	62	63
Effectif	9	8	8	6	5	4	3	3

p. 25

On suppose que le poids d'un individu du groupe est une variable aléatoire X qui suit une loi normale d'écart-type $\sigma = 3,5$. Dans chaque groupe de 100 personnes étudié, on désigne par X_i la variable aléatoire égale au poids du i-ème individu, pour tout $i \in [\![1,100]\!]$.

1. En utilisant les deux méthodes précédentes, déterminer une valeur approchée d'un intervalle de confiance à 95%, de la moyenne des poids des individus.
2. Comparer les deux méthodes.

2.2 Intervalle de confiance asymptotique

DÉFINITION — **intervalle de confiance asymptotique**

Soient $\alpha \in]0;1[$, U_n *et* V_n *deux estimateurs de* $g(\theta)$ *tels que pour tout* $\theta \in \Theta$

$$\mathbf{P}_\theta\Big(\, g(\theta) \in [U_n;V_n] \,\Big) \geqslant 1-\alpha_n \quad \textit{et} \quad \alpha_n \underset{n\to\infty}{\longrightarrow} \alpha.$$

On dit que l'intervalle $[U_n;V_n]$ *est un* ***intervalle de confiance asymptotique*** *de* $g(\theta)$ *avec un risque d'au plus* α *ou au niveau de confiance au moins égal à* $1-\alpha$.

Intervalle de confiance asymptotique du paramètre d'une loi de Bernoulli

Soit $(X_1, X_2, \ldots, X_n)$ un n-échantillon issu d'une loi de Bernoulli de paramètre p. Par indépendance, on sait que

$$n\overline{X_n} = \sum_{i=1}^{n} X_i \hookrightarrow \mathscr{B}(n,p).$$

Par le théorème central limite, on sait de plus que

$$\overline{X_n}^* = \frac{\overline{X_n} - \mathbf{E}\left(\overline{X_n}\right)}{\sqrt{\mathbf{V}\left(\overline{X_n}\right)}} = \sqrt{n}\left(\frac{\overline{X_n} - p}{\sqrt{p(1-p)}}\right) \underset{n\to+\infty}{\overset{\mathscr{L}}{\longrightarrow}} Z \quad \text{où} \quad Z \hookrightarrow \mathscr{N}(0,1).$$

Autrement dit, pour tous $a, b \in \mathbb{R}$ avec $a < b$

$$\mathbf{P}\left(a < \overline{X_n}^* \leqslant b\right) \underset{n\to\infty}{\longrightarrow} \mathbf{P}(a < Z \leqslant b) = \Phi(b) - \Phi(a).$$

En particulier, pour un intervalle centré en 0, $a = -b$ et en reprenant les notations précédentes : soit t_α tel que $\Phi(t_\alpha) = 1 - \frac{\alpha}{2}$.

$$\begin{aligned}\mathbf{P}\left(\overline{X_n} - t_\alpha\frac{\sqrt{p(1-p)}}{\sqrt{n}} \leqslant p \leqslant \overline{X_n} + t_\alpha\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right) &= \mathbf{P}\left(-t_\alpha\frac{\sqrt{p(1-p)}}{\sqrt{n}} \leqslant \overline{X_n} - p \leqslant t_\alpha\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right)\\ &= \mathbf{P}\left(-t_\alpha < \sqrt{n}\frac{\overline{X_n} - p}{\sqrt{p(1-p)}} \leqslant t_\alpha\right)\\ &= \mathbf{P}(t_\alpha < \overline{X_n}^* \leqslant t_\alpha) \underset{n\to\infty}{\longrightarrow} \Phi(t_\alpha) - \Phi(-t_\alpha) = 2\Phi(t_\alpha) - 1 = 1 - \alpha.\end{aligned}$$

Ensuite, à partir de l'encadrement $0 \leqslant p(1-p) \leqslant \frac{1}{4}$

$$\left[\overline{X_n} - t_\alpha\frac{\sqrt{p(1-p)}}{\sqrt{n}} \leqslant p \leqslant \overline{X_n} + t_\alpha\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right] \subset \left[\overline{X_n} - \frac{t_\alpha}{2\sqrt{n}} \leqslant p \leqslant \overline{X_n} + \frac{t_\alpha}{2\sqrt{n}}\right].$$

Par croissance de la probabilité

$$\mathbf{P}\left(\left[\overline{X_n} - t_\alpha\frac{\sqrt{p(1-p)}}{\sqrt{n}} \leqslant p \leqslant \overline{X_n} + t_\alpha\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right]\right) \leqslant \mathbf{P}\left(\left[\overline{X_n} - \frac{t_\alpha}{2\sqrt{n}} \leqslant p \leqslant \overline{X_n} + \frac{t_\alpha}{2\sqrt{n}}\right]\right).$$

Si, pour tout $n \in \mathbb{N}$, on pose α_n de sorte que $1-\alpha_n$ soit le terme de gauche dans l'inégalité précédente, on a bien trouvé une suite $(\alpha_n)_{n\in\mathbb{N}^*}$ telle que

$$\mathbf{P}\left(\overline{X_n} - \frac{t_\alpha}{2\sqrt{n}} \leqslant p \leqslant \overline{X_n} + \frac{t_\alpha}{2\sqrt{n}}\right) \geqslant 1 - \alpha_n \quad \text{et} \quad \alpha_n \underset{n\to\infty}{\longrightarrow} \alpha.$$

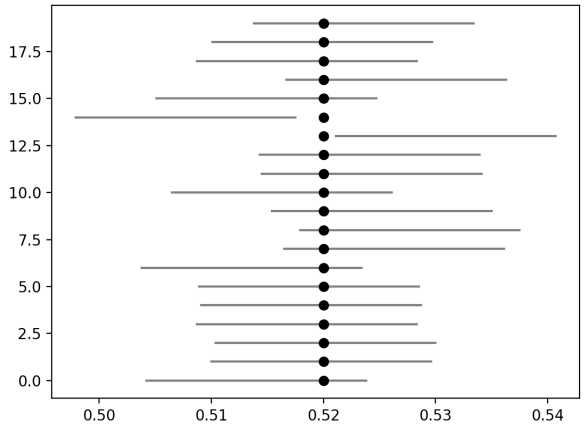
En reprenant la définition, un intervalle de confiance asymptotique de p au niveau de confiance $1-\alpha$ est

$$\left[\overline{X_n} - \frac{t_\alpha}{2\sqrt{n}}, \overline{X_n} + \frac{t_\alpha}{2\sqrt{n}}\right].$$

Simulation Python. Le code suivant crée plusieurs séries de données et construit l'intervalle aléatoire obtenu par l'intervalle de confiance asymptotique associé à chaque série de données.

Editeur

```
p=0.52 # le paramètre "inconnu" à estimer

n=10000 # Nombre de sondés
alpha=0.05; talpha=1.96

NbreTest=20
for i in range(NbreTest):
    Ech=np.random.rand(n)<p
    u=np.mean(Ech)-talpha/(2*(n)**(1/2))
    v=np.mean(Ech)+talpha/(2*(n)**(1/2))
    plt.plot([u,v],[i,i])
    plt.plot([p],[i],'o')
plt.show()
```

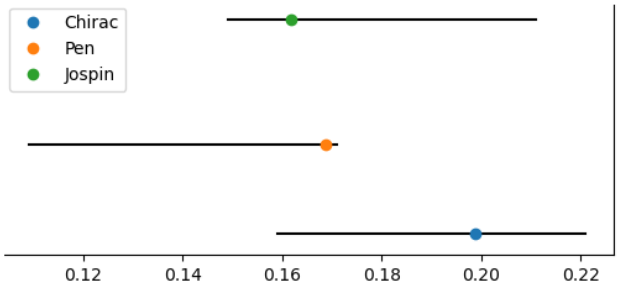


Remarque. Les intervalles obtenus sont bien plus précis que ceux obtenus par l'inégalité de Bienaymé-Tchebychev. En effet, les longueurs des intervalles obtenus par le théorème central limite sont plus petites que celles obtenues par l'inégalité de Bienaymé-Tchebychev. Par contre, on s'attend à ce que la probabilité que le paramètre ne soit pas dans l'intervalle construit à partir des données (c'est le cas ici du sixième test en partant du haut) soit plus grande.

Exemple. *Premier tour de l'élection présidentielle de 2002 (21 avril)*

Voici quelques résultats des sondages d'opinion quelques jours avant l'élection.

	Jacques CHIRAC	Jean Marie LE PEN	Lionel JOSPIN
CSA - 11 avril	21%	12%	19%
IFOP - 12 avril	19%	11,5%	17%
SOFRES - 13 avril	20%	13%	18%
CSA - 18 avril	19,15%	14%	18%
IPSOS - 18 avril	20%	14%	18%
BVA - 19 avril	**19**%	**14**%	**18**%
Sondage confidentiel - 21 avril	18%	14,5%	17%
Résultat final	**19,88**%	**16,88**%	**16,18**%



Extrait du site du conseil constitutionnel : "Dans le cas précis de l'élection présidentielle, la publication d'un sondage d'intention de vote prévoyant un faible écart entre les candidats (51-49 par exemple) rend nécessaire une telle précaution si l'on considère que, dans le cas des sondages portant sur un échantillon de 1 000 personnes (qui est l'échantillon standard en France pour les élections nationales et notamment l'élection présidentielle), la marge d'erreur est estimée à 3%.
L'obligation pour les instituts de sondage de rendre publique leur marge d'erreur est prévue par la proposition de loi sénatoriale."

Exercices

Exercice 9. ✧ **Estimation d'une variance**

Soient $n \in \mathbb{N}^*$, $p \in]0;1[$ et une variable aléatoire $X \hookrightarrow \mathscr{B}(n;p)$. Démontrer qu'il existe deux réels $\alpha_n, \beta_n \in \mathbb{N}^*$ (indépendants de p) tels que la variable aléatoire $\alpha_n X + \beta_n X^2$ soit un estimateur sans biais de $\mathbf{V}(X)$.

≫ Solution p. 25

Exercice 10. ✦ Soient $\theta \in]0,1[$ un paramètre inconnu et X une variable aléatoire de loi définie par

$$\forall\, k \in \mathbb{N}, \qquad \mathbf{P}(X=k) = (k+1)(1-\theta)^2\theta^k.$$

1. Justifier que X admet un moment d'ordre 1 et 2. Préciser l'espérance $\mathbf{E}(X)$.

   Soit $(X_1,\dots,X_n)$ un échantillon de même loi que X. On pose $\quad S_n = \dfrac{1}{n}\sum_{i=1}^{n} X_i$.

2. Expliciter une fonction bijective $g :]0;1[\to \mathbb{R}_*^+$ telle que $\mathbf{E}(S_n) = g(\theta)$.
3. En déduire que $g^{-1}(S_n)$ est un estimateur convergent de θ.

≫ Solution p. 26

Exercice 11. ✦ **Estimation d'un paramètre d'une loi uniforme**

Soient $n \in \mathbb{N}^*$ et $\theta \in \mathbb{R}_*^+$. Soient X une variable aléatoire suivant une loi uniforme sur $[0;\theta[$ et $(X_1,\dots,X_n)$ un échantillon de même loi que X.

1. Montrer que pour tout $n \in \mathbb{N}^*$, $T_n = \frac{2}{n}(X_1 + \cdots + X_n)$ est un estimateur sans biais de θ.
2. Considérons $Y_{\min}$ et $Y_{\max}$ définies par :

$$Y_{\min} = \min_{i\in[\![1;n]\!]} X_i \quad \text{et} \quad Y_{\max} = \max_{i\in[\![1;n]\!]} X_i.$$

   Pour chacune de ces deux variables, préciser la fonction de répartition, une densité de probabilité, l'espérance et la variance.
   On pourra remarquer que $Y_{\min}$ *et* $\theta - Y_{\max}$ *ont même loi.*
3. Posons $T'_n = \frac{n+1}{n} Y_{\max}$. Justifier que T'_n est un estimateur sans biais de θ.
4. Quel est le meilleur estimateur de θ entre T'_n et T_n ?
   Un estimateur sans biais est d'autant meilleur que sa variance est faible.
5. Posons $T''_n = Y_{\min} + Y_{\max}$. Montrer sans calculs superflus que $\mathbf{V}(T''_n) \leqslant 4\mathbf{V}(Y_{\max})$.
6. En déduire que T''_n est un estimateur sans biais de θ meilleur que T_n pour n suffisamment grand.
7. Est-ce que les estimateurs T_n, T'_n et T''_n sont convergents ?

≫ Solution p. 26

Exercice 12. ✦ **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 $\mathbf{P}_\theta$. Dans ce cas, on appelle risque quadratique de T_n en $g(\theta)$ et on note $r_\theta(T_n)$ le réel :

$$r_\theta(T_n) = \mathbf{E}_\theta\left(\left(T_n - g(\theta)\right)^2\right).$$

1. Justifier que $r_\theta(T_n) = \left(b_\theta(T_n)\right)^2 + \mathbf{V}_\theta(T_n)$.
2. **a)** Justifier que si le risque quadratique de T_n tend vers 0 quand n tend vers l'infini, alors T_n est un estimateur convergent de $g(\theta)$.

   b) En déduire que, si T_n est un estimateur asymptotiquement sans biais de $g(\theta)$ (c'est-à-dire $\mathbf{E}(T_n) \underset{n\to\infty}{\longrightarrow} g(\theta)$) et si la variance de T_n tend vers 0 lorsque n tend vers l'infini, alors T_n est un estimateur convergent de $g(\theta)$.
3. Pour comparer deux estimateurs T_n et U_n de $g(\theta)$, on peut calculer leur risque quadratique. Si on a :

$$\forall\, \theta \in \Theta, \qquad r_\theta(T_n) \leqslant r_\theta(U_n),$$

   alors on dira que T_n est un meilleur estimateur de $g(\theta)$ que U_n.

a) *Exemple 1.*
Soit $(X_n)_{n\in\mathbb{N}^*}$ une suite de variables aléatoires indépendantes suivant la loi exponentielle $\mathscr{E}(\lambda)$. 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 ces deux estimateurs de $\theta = 1/\lambda$.

b) *Exemple 2.*
Soient $T_{n,0}$ et $T_{n,1}$ deux estimateurs de θ, sans biais et indépendants. Pour tout a réel, on pose

$$T_{n,a} = aT_{n,1} + (1-a)T_{n,0}.$$

i) Est-ce que $T_{n,a}$ est un estimateur sans biais de θ ?

ii) Pour quelle valeur de a, l'estimateur est-il le meilleur ?

≫ Solution p. 27

Exercice 13. ✦✦ **Estimateur sans biais de l'écart-type σ d'une loi normale centrée** *extrait HEC 2005*

Soit X une variable aléatoire qui suit une loi normale centrée et d'écart-type σ, le paramètre réel inconnu σ, est strictement positif.

1. Montrer que la variable aléatoire $T = \frac{X^2}{2\sigma^2}$ suit une loi γ de paramètre $1/2$. En déduire la valeur de $\Gamma(1/2)$.

2. Pour n entier naturel non nul, on considère un n-échantillon $(X_1, X_2, \ldots, X_n)$ constitué de variables indépendantes et de même loi que X.

a) On désigne par S_n la variable aléatoire $S_n = \sum_{i=1}^{n} \frac{{X_i}^2}{2\sigma^2}$. Quelle est la loi de probabilité de S_n ?

b) En déduire que la variable aléatoire Y_n définie par $Y_n = \frac{1}{n}\sum_{i=1}^{n} {X_i}^2$ est un estimateur sans biais de σ^2.

≫ Solution p. 32

Exercice 14. ✦ **Estimation et optimisation sous contrainte**

Soient $n \in \mathbb{N}^*$ et $\theta \in \mathbb{R}_*^+$. Soient X une variable aléatoire d'espérance $\mathbf{E}(X) = \theta \neq 0$, de variance $\mathbf{V}(X) = 1$ et $(X_1, \ldots, X_n)$ un échantillon de même loi que X. On pose $\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

1. Montrer que $\overline{X_n}$ est un estimateur sans biais de θ et préciser sa variance.

2. Soit $(\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^n$. On note $Y_n = \sum_{i=1}^{n} \alpha_i X_i$.

a) Donner une condition nécessaire et suffisante sur $\alpha_1, \ldots, \alpha_n$ pour que Y_n soit un estimateur sans biais de θ. On suppose dans la suite que cette condition est vérifiée.

b) Calculer $\mathrm{Cov}\left(\overline{X_n}, Y_n\right)$. En déduire $\mathbf{V}\left(\overline{X_n}\right) \leqslant \mathbf{V}(Y_n)$.

c) Donner un choix de paramètres $\alpha_1, \ldots, \alpha_n$ afin que la variance de Y_n soit minimale ?

≫ Solution p. 28

Exercice 15. ✦✦ **Loi exponentielle translatée**

Soient $\theta \in \mathbb{R}$ et Y une variable aléatoire suivant une loi exponentielle de paramètre 1. On considère la variable aléatoire $X = Y + \theta$ de densité f_θ définie par :

$$\forall x \in \mathbb{R}, \qquad f_\theta(x) = \exp\big(-(x-\theta)\big)\mathbf{1}_{[\theta,+\infty[}(x).$$

1. Préciser l'espérance et la variance de X.

2. Soit $(X_1, \ldots, X_n)$ un échantillon de même loi que X. Vérifier que $T_n = \overline{X_n} - 1$ est un estimateur convergent de θ.

3. *Estimation par le maximum de vraisemblance*
Soient $x_1, \ldots, x_n$ des réels fixés. On définit la fonction L sur $\mathbb{R}$ par

$$\forall \theta \in \mathbb{R}, \qquad L(\theta) = \prod_{k=1}^{n} f_\theta(x_k).$$

a) i) Que de dire de $L(\theta)$ si l'un des réels x_k est inférieur strictement à θ ?

ii) Dans le cas où $\theta \leqslant \min(x_1, \ldots, x_n)$, expliciter $\ln\big(L(\theta)\big)$.

iii) En déduire que la fonction L admet un maximum atteint uniquement en $\widehat{\theta} = \min(x_1, \ldots, x_n)$.

b) Ce résultat justifie le choix de considérer l'estimateur : $S_n = \min(X_1, \ldots, X_n)$.
On a vu que si $(Y_1, \ldots, Y_n)$ est un échantillon de même loi que Y alors $\min(Y_1, \ldots, Y_n)$ suit une loi exponentielle $\mathscr{E}(n)$. En déduire la loi de S_n. Justifier que S_n est un estimateur convergent de θ.

c) Quel est le meilleur estimateur entre T_n et S_n ?
On regardera quel estimateur à la risque quadratique $r_\theta(T_n) = \left(b_\theta(T_n)\right)^2 + \mathbf{V}_\theta(T_n)$ *le plus petit.*

4. *Intervalle de confiance*
Soit $\alpha \in]0;1[$. On pose $\gamma_{n,\alpha} = |\ln(\alpha)|/n$. Montrer que

$$\mathbf{P}(S_n - \gamma_{n,\alpha} \leqslant \theta \leqslant S_n) = 1 - \alpha.$$

≫ Solution p. 28

Exercice 16. ✦✦ *d'après EDHEC 2000*

Un sondage consiste à proposer l'affirmation « A » à certaines personnes d'une population donnée. Le sujet abordé étant délicat, le stratagème suivant est mis en place afin de mettre en confiance les personnes sondées pour qu'elles ne mentent pas...
L'enquêteur dispose d'un paquet de 20 cartes, numérotées de 1 à 20 , qu'il remet à la personne sondée. Celle-ci tire une carte au hasard et ne la montre pas à l'enquêteur. La règle est alors la suivante :

- si la carte porte le numéro 1, la personne sondée répond "vrai" si elle est d'accord avec l'affirmation « A » et "faux" sinon.
- si la carte porte un autre numéro, la personne sondée répond "vrai" si elle n'est pas d'accord avec l'affirmation « A » et "faux" sinon.

Le but de l'enquête est d'évaluer la proportion p de gens de cette population qui sont réellement d'accord avec l'affirmation « A ».

1. On interroge une personne selon ce procédé et on considère l'événement suivant, noté V : «la personne répond "vrai"». On note $\theta = \mathbf{P}(V)$. En utilisant la formule des probabilités totales, exprimer θ en fonction de p, puis en déduire p en fonction de θ.

2. Certaines considérations théoriques laissent penser que $p = \frac{17}{18}$.

a) Vérifier que $\theta = \frac{1}{10}$.

b) Calculer la probabilité pour qu'une personne ayant répondu "vrai" soit d'accord avec l'affirmation « A».

On revient au cas général où l'on ne connaît ni p, ni θ.

3. On considère un échantillon aléatoire, de taille n, extrait de la population considérée et on note S_n le nombre de réponses "vrai" obtenues. On suppose n assez grand pour pouvoir considérer que cet échantillonnage est assimilable à un tirage avec remise.

a) Donner la loi de S_n ainsi que son espérance et sa variance.

b) Montrer que $\frac{S_n}{n}$ est un estimateur sans biais et convergent de θ.

4. Dans cette question, on suppose que l'on a réalisé un échantillon de 100 personnes et on constate que 23 personnes ont répondu "vrai".

a) Donner une estimation ponctuelle de θ et de p.

b) Donner un intervalle de confiance à 95% de θ puis de p.
On rappelle que, si Φ désigne la fonction de répartition d'une variable X suivant la loi normale $\mathcal{N}(0,1)$, alors $\Phi(1,96) = 0,975$

≫ Solution p. ??

Exercice 17. ✦✦ **Estimation des paramètres d'une loi de Pareto**

Soient $(b,\theta) \in \mathbb{R}^{+*} \times \mathbb{R}$. On suppose que X est une va à densité qui prend ses valeurs dans $[\theta;+\infty[$, dont une densité f est définie par :

$$\forall x \in \mathbb{R}, \quad f(x) = \begin{cases} 0 & \text{si } x < \theta \\ \frac{1}{b}\exp\left(-\frac{x-\theta}{b}\right) & \text{si } x \geqslant \theta. \end{cases}$$

Soient $n \in \mathbb{N} \setminus \{0;1\}$ et $(X_1,\ldots,X_n)$ un n-échantillon de même loi que X. On pose :

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

1. Soit $k \in [\![1;n]\!]$. Reconnaitre la loi de $Y_k = (X_k - \theta)/b$.

2. On pose $\overline{Y_n} = \frac{Y_1+\cdots+Y_n}{n}$ et $U_n = \min(Y_1,\ldots,Y_n)$. Calculer $\mathbf{E}\left(\overline{Y_n}\right)$ et $\mathbf{E}(U_n)$.

3. Exprimer les variables aléatoires $\overline{X_n}$ et T_n en fonction des variables $\overline{Y_n}$ et U_n. En déduire $\mathbf{E}\left(\overline{X_n}\right)$ et $\mathbf{E}(T_n)$.

4. Déterminer un estimateur sans biais $\widehat{\theta}_n$ de θ et un estimateur sans biais $\widehat{b}_n$ de b sous la forme de combinaisons linéaires de $\overline{X_n}$ et T_n.

≫ Solution p. 29

Exercice 18. ✦✦ **Estimations des paramètres d'une loi de Pareto**

Une variable X suit une loi de Pareto VP(α, 1, x_0) si la fonction de répartition est donnée par

$$\forall\, x \in \mathbb{R}, \qquad \mathrm{F}(x) = \begin{cases} 0 & \text{si } x < 1 + x_0 \\ 1 - \frac{1}{(x-x_0)^{\alpha}} & \text{sinon.} \end{cases}$$

1. Dans la suite, on souhaite estimer le paramètre α d'une loi VP(α, 1, 0).
   a) Donner la loi suivie par Z = ln(X).
   b) Soient $(\mathrm{X}_n)_{n\in\mathbb{N}^*}$ une suite de variables mutuellement indépendantes qui suivent toutes une loi de Pareto VP(α, 1, 0). Pour tout $n \in \mathbb{N}^*$, on pose
   $$\mathrm{Z}_n = \ln(\mathrm{X}_1\mathrm{X}_2\ldots\mathrm{X}_n).$$
   Donner la loi de Z_n.
   c) Préciser l'espérance et la variance de Z_n, en déduire que $\mathrm{T}_n = \frac{1}{n}\mathrm{Z}_n$ est un estimateur de α^{-1}.
   d) Démontrer que $\mathrm{W}_n = \frac{n}{\mathrm{Z}_n}$ est un estimateur de α, calculer son espérance, sa variance.
2. On revient au cas général. On veut maintenant estimer le paramètre x_0 d'une loi de Pareto de paramètre (α, 1, x_0).
   a) Démontrer que X suit une loi de Pareto VP(α, 1, x_0) si et seulement si $\mathrm{X}_0 = \mathrm{X} - x_0$ suit un loi de Pareto VP(α, 1, 0).
   b) Calculer $\mathbf{E}(\mathrm{X}_0)$ et $\mathbf{V}(\mathrm{X}_0)$ pour α > 2, et en déduire
   $$\mathbf{E}(\mathrm{X}) = x_0 + \frac{\alpha}{\alpha-1} \quad \text{et} \quad \mathbf{V}(\mathrm{X}) = \frac{\alpha}{(\alpha-2)(\alpha-1)^2}.$$
   c) Soient $(\mathrm{X}_n)_{n\in\mathbb{N}^*}$ une suite de variables mutuellement indépendantes qui suivent toutes une loi de Pareto VP(α, 1, x_0), on pose pour tout $n \in \mathbb{N}^*$,
   $$\mathrm{Y}_n = \inf(\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_n).$$
   Préciser la loi, l'espérance la variance de Y_n. En déduire que $\tilde{\mathrm{Y}}_n = \mathrm{Y}_n - 1$ est un estimateur de rang n de x_0. Quel est le risque lié à l'estimateur $\tilde{\mathrm{Y}}_n$?

Exercice 19. ✦✦ **Le boulanger** *d'après oral ESCP 2022, sujet 23*

Un boulanger vend du pain chaque jour.

- La quantité de pain produite chaque jour est une quantité fixée Q choisie par le boulanger, Q étant exprimée en kilogramme.
- La demande de pain de la part des clients est une variable aléatoire X strictement positive, toujours exprimée en kilogramme.
- On suppose que la variable X admet une densité f strictement positive sur $\mathbb{R}_+^*$, nulle sur $\mathbb{R}_-$, continue sur $\mathbb{R}$ et on note F la fonction de répartition de X.
- Le coût de fabrication par kilogramme est c euros et le prix de vente est v euros par kilogramme.
- On note B la variable aléatoire égale au bénéfice quotidien.
- La variable indicatrice d'un événement A est notée $\mathbb{1}_{\mathrm{A}}$.
- On suppose que $0 < c < v$.

Si la demande de pain X est inférieure à l'offre Q, le boulanger ne vend que la quantité X (le pain invendu un jour donné n'est pas remis en vente le lendemain !) ; si la demande est supérieure à l'offre, il ne vend que la quantité Q. Dans ces conditions, on cherche la quantité optimale à produire, c'est-à-dire la quantité Q_0 qui maximise l'espérance de B.

1. Établir la relation suivante :
   $$\mathrm{B} = v\left[\mathrm{Q} + (\mathrm{X} - \mathrm{Q})\mathbf{1}_{[\mathrm{X}<\mathrm{Q}]}\right] - c\mathrm{Q}.$$
2. Montrer que la variable $\mathrm{X}\mathbf{1}_{[\mathrm{X}<\mathrm{Q}]}$ admet une espérance et donner son expression sous forme d'intégrale.
3. En déduire l'égalité suivante :
   $$\mathbf{E}(\mathrm{B}) = (v-c)\mathrm{Q} + v\left[\int_0^{\mathrm{Q}} t f(t)\,\mathrm{d}t - \mathrm{QF}(\mathrm{Q})\right].$$
4. Exprimer Q_0 à l'aide de F, de v et de c. Le boulanger cherche à prévoir sa demande journalière. La demande aléatoire X_n qui va s'exprimer le jour n n'est pas connu à l'avance mais le boulanger fait l'hypothèse que la demande ne variera pas beaucoup d'un jour à l'autre et que :
   $$\mathrm{X}_{n+1} = \mathrm{X}_n + \mathrm{U}_{n+1}$$
   où :
   - X_0 est une constante strictement positive fixée.
   - Les U_k sont des variables aléatoires indépendantes, de même loi, d'espérance nulle et de variance σ^2 non nulle.

5. **a)** Exprimer X_n en fonction des U_i et de X_0.

b) Montrer que la suite $\left(\frac{X_0}{n}\right)$ converge en probabilité vers 0.

c) Démontrer que si deux suites de variables aléatoires (A_n) et (B_n) convergent en probabilité respectivement vers des variables aléatoires a et b, alors la suite $(A_n + B_n)$ converge en probabilité vers $a + b$.

d) En déduire que $\left(\frac{X_n}{n}\right)$ converge en probabilité vers une variable que l'on précisera.

e) Montrer que la suite $\left(\frac{X_0}{n}\right)$ converge en loi vers 0.

f) Montrer que la suite $\left(\frac{X_n}{n}\right)$ converge en loi et préciser la loi limite.
On pourra utiliser le Théorème de Slutsky : Si X_n converge en loi vers X, et si Y_n converge en probabilité vers une constante c, alors $X_n Y_n$ converge en loi vers cX.

≫ Solution p. ??

Exercice 20. ✦✦ **Calcul d'un intervalle de confiance**

Soit $(X_1, \dots, X_n)$ un échantillon de la loi exponentielle de paramètre $\lambda > 0$ inconnu. On pose

$$Y_n = \min(X_1, \dots, X_n).$$

1. Montrer que Y_n suit une loi exponentielle dont on précisera le paramètre.

2. Soit $\alpha \in]0;2/3[$. Déterminer deux réels a_n et b_n tels que

$$\mathbf{P}(n\lambda Y_n \leqslant a_n) = \frac{\alpha}{2} = \mathbf{P}(n\lambda Y_n \geqslant b_n).$$

3. Justifier que $\left[\frac{nY_n}{a_n}; \frac{nY_n}{b_n}\right]$ est un intervalle de confiance de $\frac{1}{\lambda}$ au niveau de confiance d'au moins $1-\alpha$.

≫ Solution p. 33

Exercice 21. ✦✦ **Calcul d'un intervalle de confiance II**

Soit $(X_i)_{i\in\mathbb{N}^*}$ une suite de variables aléatoires indépendantes et de même loi exponentielle $\mathscr{E}(\lambda)$ où le paramètre $\lambda > 0$ est inconnu. On pose $\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

1. Justifier que $\lambda\sqrt{n}\overline{X_n} - \sqrt{n}$ converge en loi vers une variable aléatoire de loi $\mathscr{N}(0,1)$.

2. Soit $\alpha \in]0,1[$. Notons t_α l'unique réel tel que $\Phi(t_\alpha) = 1 - \frac{\alpha}{2}$. Montrer que

$$\left[\left(1 - \frac{t_\alpha}{\sqrt{n}}\right)\frac{1}{\overline{X_n}}, \left(1 + \frac{t_\alpha}{\sqrt{n}}\right)\frac{1}{\overline{X_n}}\right]$$

est un intervalle de confiance asymptotique de λ au niveau de risque α.

≫ Solution p. ??

Exercice 22. ✦✦ **Estimation et loi de Poisson** *d'après oral ESCP 2022, sujet 35*

Soit $X_1, \dots, X_n$ un n-échantillon de variables aléatoires indépendantes qui suivent la loi de Poisson de paramètre $\lambda \in \mathbb{R}_*^+$ inconnu.
On cherche dans cet exercice à estimer $e^{-\lambda}$.
Pour tout $k \in [\![1, n]\!]$, on note Y_k la fonction indicatrice de l'événement $[X_k = 0]$. Pour tout $n \in \mathbb{N}^*$, on pose :

$$\overline{Y_n} = \frac{1}{n}\sum_{k=1}^{n} Y_k \quad \text{et} \quad S_n = \sum_{k=1}^{n} X_k$$

1. **a)** Déterminer la loi de Y_k.

b) Montrer que $\overline{Y_n}$ est un estimateur sans biais de $e^{-\lambda}$.

c) $\overline{Y_n}$ est-il un estimateur convergent de $e^{-\lambda}$?

2. Pour $j \in \mathbb{N}$, on pose $\varphi(j) = \mathbf{P}_{[S_n = j]}(X_1 = 0)$. Calculer $\varphi(j)$.

3. On pose à présent $T_n = \varphi(S_n)$.

a) Montrer que T_n est un estimateur sans biais de $e^{-\lambda}$.

b) T_n est-il un estimateur convergent de $e^{-\lambda}$?

4. Comparer les variances des deux estimateurs T_n et $\overline{Y}_n$.

≫ Solution p. ??

Lien entre estimation et optimisation

Exercice 23. ✦

Soit $(X_1, X_2, \dots, X_n, Y_1, Y_2, \dots, Y_m)$ un échantillon de variables aléatoires indépendantes, de même loi de Bernoulli de paramètre inconnu $p \in]0;1[$. On pose :

$$\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \overline{Y_m} = \frac{1}{m}\sum_{i=1}^{m} Y_i \quad \text{et} \quad Z = a\overline{X_n} + b\overline{Y_m}, \quad \text{où} \quad (a,b) \in \mathbb{R}^2.$$

Pour quelles valeurs de a et b, Z est-il le meilleur estimateur sans biais de p ?

≫ Solution p. 29

Exercice 24. ✦✦ Soit X une variable aléatoire discrète d'espérance $\mathbf{E}(X) = \theta \neq 0$ et de variance $\mathbf{V}(X) = 1$. On considère un n-échantillon $(X_1, \dots, X_n)$ de la loi de X. On pose classiquement

$$\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

1. Montrer que $\overline{X_n}$ est un estimateur sans biais de θ et calculer son risque quadratique défini par

$$r_\theta = \mathbf{E}_\theta\left((\overline{X_n} - \theta)^2\right).$$

2. Soit $(\alpha_1, \dots, \alpha_n) \in \mathbb{R}^n$. On note $Y_n = \sum_{i=1}^{n} \alpha_i X_i$.

a) Donner une condition nécessaire et suffisante sur $\alpha_1, \dots, \alpha_n$ pour que Y_n soit un estimateur sans biais de θ. On suppose dans la suite que cette condition est vérifiée.

b) Calculer $\operatorname{cov}\left(\overline{X_n}, Y_n\right)$. En déduire $\mathbf{V}\left(\overline{X_n}\right) \leqslant \mathbf{V}(Y_n)$. Que dire si $\mathbf{V}\left(\overline{X_n}\right) = \mathbf{V}(Y_n)$?

c) Que peut-on en déduire sur les estimateurs $\overline{X_n}$ et Y_n de θ.

≫ Solution p. 29

Exercice 25. ✦ **Maximum de vraisemblance - cas discret**

Soit $\theta \in \mathbb{R}_*^+$. Pour tout k de $\mathbb{N}$, on pose : $\quad p_k = \alpha\left(\dfrac{\theta}{1+\theta}\right)^k \quad$ où $\quad \alpha \in \mathbb{R}$.

1. Expliciter α de sorte que $(p_k)_{k\in\mathbb{N}}$ définisse une loi de probabilité.
Dans la suite, on considère une variable aléatoire X dont la loi est donnée par

$$X(\Omega) = \mathbb{N} \quad \text{et} \quad \forall k \in \mathbb{N}, \quad \mathbf{P}(X = k) = p_k.$$

2. Reconnaître la loi de $X + 1$. En déduire l'espérance et la variance de X.

3. *Estimation de* θ *par maximum de vraisemblance*
Soient $(X_1, X_2, \dots, X_n)$ un échantillon de même loi que X et $x_1, x_2, \dots, x_n$ appartenant à $X(\Omega)$. On définit les fonctions L et ℓ par

$$\forall \theta \in \mathbb{R}_+^*, \quad L(\theta) = \prod_{k=1}^{n} \mathbf{P}\left(X_k = x_k\right) \quad \text{et} \quad \ell(\theta) = \ln\left(L(\theta)\right).$$

L'objectif est de choisir la valeur de θ où est atteint le maximum (s'il existe) de L.

a) Exprimer $\ell(\theta)$ à l'aide de $s_n = \sum_{k=1}^{n} x_k$. En déduire que ℓ admet un maximum.

b) Justifier que L admet un maximum atteint en un unique point que l'on exprimera sous la forme $\theta = f(x_1, \dots, x_n)$ où f est une fonction de $\mathbb{R}^n \to \mathbb{R}$.

4. On pose maintenant $T_n = f(X_1, \dots, X_n)$. Vérifier que T_n est un estimateur convergent de θ.

≫ Solution p. 30

Python

Exercice 26. ✦ **Retour sur la méthode de Monte-Carlo**

Soient $(U_i)_{i\in\mathbb{N}}$ des variables aléatoires réelles indépendantes de loi uniforme sur $[0,1]$ et N une variable aléatoire de loi géométrique de paramètre p indépendante de la suite $(U_i)_{i\in\mathbb{N}}$. On pose

$$X = \max_{1 \leqslant i \leqslant N} U_i.$$

1. Déterminer la fonction de répartition de la variable aléatoire X.

2. Calculer l'espérance de X.
3. Simuler la variable et vérifier votre résultat.

≫ Solution p. 30

Exercice 27. ✦ Soit $(X_1,\dots,X_n)$ un échantillon suivant la loi $\mathscr{E}(\lambda)$. Nous disposons des deux estimateurs de $\theta = \dfrac{1}{\lambda}$:

- La moyenne empirique $\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$;
- L'écart-type empirique $\sigma_n = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}\left(X_k - \overline{X_n}\right)^2}$.

On souhaite comparer ces deux estimateurs.

1. Écrire un programme qui prend en argument n et λ et simule $\overline{X_n}$, puis σ_n.
2. Afficher des histogrammes et comparer les estimateurs.

≫ Solution p. 32

Exercice 28. ✦

Soient $n \in \mathbb{N}^*$ et $(X_1, X_2, \dots, X_n)$ un échantillon d'une loi de Poisson $\mathscr{P}(\theta)$. On cherche à estimer $\exp(-\theta)$. Pour cela, on pose :

$$A_n = \left(1 - \frac{1}{n}\right)^{\sum_{i=1}^{n} X_i} \quad \text{et} \quad B_n = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{[X_i=0]}.$$

1. Écrire deux programmes qui prend en arguments n, θ et simulent respectivement A_n et B_n.
2. On suppose que $\theta = 1$.
   En déduire un programme qui prend en argument n et renvoie une approximation de l'espérance de A_n et B_n.
   Que peut-on conjecturer sur le biais de chacun de ces estimateurs?
3. Afficher l'histogramme de 10000 réalisations de T_{100} et U_{100}.
   Que peut-en déduire sur la qualité de ces estimateurs?

≫ Solution p. ??

Estimation par intervalle de confiance

Exercice 29. ✦ **Estimation de la variance d'une loi normale centrée**

Soit X une variable aléatoire qui suit une loi normale centrée (d'espérance nulle) et d'écart-type σ. L'objectif est d'estimer σ^2.

1. Montrer que la variable aléatoire $T = \frac{X^2}{2\sigma^2}$ suit une loi γ de paramètre $1/2$.
2. *Étude d'un estimateur*
   Pour $n \in \mathbb{N}^*$, on considère un échantillon $(X_1, X_2, \dots, X_n)$ de même loi que X.
   a) On désigne par S_n la variable aléatoire $S_n = \sum_{i=1}^{n} \frac{{X_i}^2}{2\sigma^2}$. Quelle est la loi de S_n ?
   b) Retrouver le fait que la variable aléatoire Y_n définie par $Y_n = \frac{1}{n}\sum_{i=1}^{n} {X_i}^2$ est un estimateur sans biais de σ^2.
   c) Est-il convergent?
3. *Calcul d'un intervalle de confiance*
   a) Soit $\alpha \in]0;1[$. Justifier qu'il existe un unique réel $k_{\alpha,n}$ tel que $\mathbf{P}(S_n \geqslant k_{\alpha,n}) = 1-\alpha$.
   b) En déduire que l'intervalle $\left[0; nY_n/(2k_{\alpha,n})\right]$ est un intervalle de confiance de σ^2 de risque α.

≫ Solution p. 32

Exercice 30. ✦✦ **Intervalle de confiance sur l'étendue d'une loi uniforme**

Soient $n \in \mathbb{N}^*$ et $\theta \in \mathbb{R}_*^+$. Soient X une variable aléatoire suivant une loi uniforme sur $[0;\theta[$ et $(X_1,\dots,X_n)$ un échantillon de même loi que X. On pose

$$Y_n = \max_{i \in [\![1;n]\!]} X_i.$$

1. Justifier que la suite $\left(n(1-Y_n/\theta)\right)_n$ converge en loi vers une variable Y de loi exponentielle $\mathscr{E}(1)$.
2. a) Soit F la fonction de répartition de Y. Justifier que pour tout $\alpha \in]0;1[$, tout $n \in \mathbb{N}^*$, il existe un réel γ_n inférieur à 1 tel que
   $$F\left(n(1-\gamma_n)\right) = 1-\alpha.$$

b) En déduire que $\left[Y_n ; \dfrac{Y_n}{\gamma_n}\right]$ est un intervalle asymptotique de θ au niveau de risque α.

≫ Solution p. 33

Exercice 31. ✦✦ **Calcul d'un intervalle de confiance pour une loi exponentielle**

Soit $(X_1, \ldots, X_n)$ un échantillon de la loi exponentielle de paramètre $\lambda \in \mathbb{R}_*^+$. L'objectif est d'estimer $1/\lambda$. On pose

$$\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{et} \quad T_n = \min(X_1, \ldots, X_n).$$

1. **a)** Reconnaître la loi de T_n.

b) Vérifier que nT_n et $\overline{X_n}$ sont deux estimateurs sans biais de $1/\lambda$.

2. *Calcul d'un intervalle de confiance avec* T_n

a) Déterminer deux réels a et b tels que $\mathbf{P}(n\lambda T_n \leqslant a) = \dfrac{\alpha}{2} = \mathbf{P}(n\lambda T_n \geqslant b)$.

b) Justifier que $\left[\frac{nT_n}{b} ; \frac{nT_n}{a}\right]$ est un intervalle de confiance de $\frac{1}{\lambda}$ au niveau de confiance de risque au plus α. Préciser l'étendue moyenne de cet intervalle.

3. *Calcul d'un intervalle de confiance avec* $\overline{X_n}$

a) Justifier que la suite $\left(\lambda\sqrt{n}\overline{X_n} - \sqrt{n}\right)_n$ converge en loi vers une variable aléatoire de loi $\mathcal{N}(0;1)$.

b) Soit $\alpha \in]0;1[$. Notons t_α l'unique réel tel que $\Phi(t_\alpha) = 1 - \frac{\alpha}{2}$. Montrer que

$$\left[\left(1 - \frac{t_\alpha}{\sqrt{n}}\right)\frac{1}{\overline{X_n}}, \left(1 + \frac{t_\alpha}{\sqrt{n}}\right)\frac{1}{\overline{X_n}}\right]$$

est un intervalle de confiance asymptotique de λ au niveau de risque α.

≫ Solution p. 33