CHAPITRE

XIII

ESTIMATION

Sommaire

A - Introduction au concept d'estimation

A.1 - Premiers exemples

L'objectif de ce chapitre est d'introduire le vocabulaire et la démarche de la statistique inférentielle en abordant, sur quelques cas simples, le problème de l'estimation, ponctuelle ou par intervalle de confiance. On se restreindra à une famille de lois de probabilités indexées par un paramètre scalaire (ou vectoriel) dont la valeur (scalaire ou vectorielle) caractérise la loi. On cherche alors à estimer la valeur du paramètre (ou une fonction simple de ce paramètre) à partir des données disponibles.

Dans ce contexte, on considère un phénomène aléatoire et on s'intéresse à une variable aléatoire réelle X qui lui est liée, dont on suppose que la loi de probabilité n'est pas complètement spécifiée et appartient à une famille de lois dépendant d'un paramètre θ décrivant un sous-ensemble Θ de $\mathbb{R}$ (éventuellement de $\mathbb{R}^2$). 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 consiste alors à estimer la vraie valeur du paramètre θ ou de $g(\theta)$ (fonction à valeurs réelles du paramètre θ), 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...

On supposera que cet échantillon est la réalisation de n variables aléatoires $X_1, \dots, X_n$ définies sur un même espace probabilisable $(\Omega, \mathcal{A})$ muni d'une famille de probabilités $(\mathbb{P}_\theta)_{\theta\in\Theta}$. Les $X_1, \dots, X_n$ seront supposées $\mathbb{P}_\theta$-indépendantes et de même loi que X pour tout θ.

Exemples

1 ▹ Voici le nombre de buts marqués en Ligue 1 de Football, 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** et on **modélise** la situation en supposant que le nombre de buts lors d'une journée est une variable aléatoire X qui suit une loi de Poisson $\mathscr{P}(\lambda)$.

Pour **estimer** λ, on peut exploiter le fait que X admet une espérance $\mathbb{E}(X) = \lambda$ donc considérer la valeur moyenne du nombre de buts (ici $\lambda \approx 28,1$).

On peut alors **tester** le modèle considéré et faire de la prédiction.

2 ▹ On s'intéresse à la durée de vie d'un composant électrique utilisé «normalement» et sans en considérer l'usure ce qui suggère de modéliser sa durée de vie X par une loi exponentielle $\mathscr{E}(\lambda)$ (loi sans mémoire).

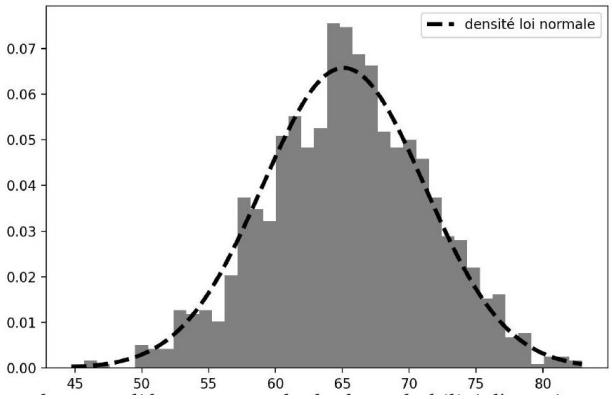
La demi-vie de X est le réel t tel que $\mathbb{P}(X \leqslant t) = \mathbb{P}(X \geqslant t)$ c'est-à-dire :

$$1 - e^{-\lambda t} = e^{-\lambda t} \quad i.e. \quad e^{-\lambda t} = \frac{1}{2} \quad \text{puis} \quad t = \frac{\ln(2)}{\lambda}.$$

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 $1000 = \ln(2)/\lambda$ *i.e.* $\lambda = \ln(2)/1000$.

À partir de ce modèle, on s'attend à ce que la durée de vie moyenne d'un composant soit $\mathbb{E}(X) = \frac{1}{\lambda}$, soit environ 1500 heures.

3 ▹ Un producteur d'œufs analyse sa production annuelle et constate la répartition suivante (histogramme des différents poids en grammes) :



Cela suggère de modéliser la situation en supposant que le poids d'un œuf est une variable aléatoire de loi normale $\mathcal{N}(\mu, \sigma^2)$ où l'on constate que $\mu \approx 65,1$ et $\sigma \approx 6,1$ (rappelons qu'en Python, la moyenne et l'écart-type s'obtiennent à l'aide des fonctions `np.mean` et `np.std`).

À l'aide de ce modèle, on peut calculer la probabilité qu'un œuf ait un poids supérieur à 73 grammes :

$$\mathbb{P}((X > 73) = 1 - \mathbb{P}(X \leqslant 73) = 1 - \mathbb{P}\left(\frac{X - \mu}{\sigma} \leqslant \frac{73 - 65,1}{6,07}\right)$$

ce qui donne :

$$\mathbb{P}((X > 73) \approx 1 - \Phi(1,30) \approx 0,0968.$$

A.2 - Notion d'estimateur

Dasn ce qui suit, on considère :

- un espace probabilisable $(\Omega, \mathscr{A})$;
- une partie Θ de $\mathbb{R}^n$ appelée espace des paramètres et telle que, pour chaque paramètre $\theta \in \Theta$, il existe une probabilité $\mathbb{P}_\theta$ sur $(\Omega, \mathscr{A})$;
- une application X qui soit une variable aléatoire sur les espaces probabilisés $(\Omega, \mathscr{A}, \mathbb{P}_\theta)$;
- sous réserve d'existence, on note $\mathbb{E}_\theta(X)$ et $\mathbb{V}_\theta(X)$ l'espérance et la variance de X pour $\mathbb{P}_\theta$.

Exemples

1 ▹ Dans l'exemple des buts marqués, on a $\Theta =]0, +\infty[$ et :

$$\forall \theta > 0, \forall k \in \mathbb{N},\ \mathbb{P}_\theta(k) = e^{-\theta}\frac{\theta^k}{k!}.$$

2 ▹ De même pour l'exemple de la durée de vie du composant : $\Theta =]0, +\infty[$.

3 ▹ Dans l'exemple de la taille des œufs, on note connait ni l'espérance, ni la variance de la loi donc $\theta = (\mu, \sigma)$ et $\Theta = \mathbb{R} \times]0, +\infty[$.

Définition XIII-1

Soit $X : \Omega \to \mathbb{R}$ une variable aléatoire définie sur $(\Omega, \mathscr{A})$ et $n \in \mathbb{N}^*$.

On appelle ***n*-échantillon** de la loi de X toute famille $(X_1, \ldots, X_n)$ de variables aléatoires sur $(\Omega, \mathscr{A})$ telles que, pour tout $\theta \in \Theta$, $X_1, \ldots, X_n$ soient $\mathbb{P}_\theta$-indépendantes et de même loi que X.

On dit aussi que la loi de X est la loi parente (ou encore loi mère) de l'échantillon.

On note «$(X_1,\dots,X_n)$ i.i.d» pour signaler que les variables sont indépendantes, et identiquement distribuées (*i.e.* de même loi).

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

Définition XIII-2

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

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

Notons qu'un estimateur ne dépend pas de θ puisque c'est la valeur que l'on souhaite déterminer.

Estimer ponctuellement $g(\theta)$ par $\varphi(x_1,\dots,x_n)$ où $\varphi(X_1,X_2,\dots,X_n)$ est un estimateur de $g(\theta)$ et $(x_1,\dots,x_n)$ est une réalisation de l'échantillon $(X_1,\dots,X_n)$, c'est décider d'accorder à $g(\theta)$ la valeur $\varphi(x_1,\dots,x_n)$.

Exemple

Avec les notations de la définition, soit $(X_1,\dots,X_n)$ un n-échantillon de la loi de X . L'***estimateur de la moyenne empirique*** est donné par :

$$\overline{X_n}=\frac{X_1+X_2+\cdots+X_n}{n}.$$

A.3 - Comparaison des estimateurs

Définition XIII-3

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

On définit le ***biais*** de T_n en $g(\theta)$ par :

$$b_\theta(T_n)=\mathbb{E}_\theta(T_n)-g(\theta).$$

Si, pour tout $\theta\in\Theta$, $b_\theta(T_n)$ est nul, alors on dit que l'estimateur est ***sans biais***, sinon on dit que l'estimateur est ***biaisé***.

Exemples

1 ▹ Considérons l'exemple de la moyenne empirique (dans le cas où X admet une espérance θ) :

$$\mathbb{E}_\theta(\overline{X_n})=\frac{1}{n}\sum_{k=1}^{n}\underbrace{\mathbb{E}_\theta(X_k)}_{=\theta}=\theta$$

donc $b_\theta(\overline{X_n})=0$ et l'estimateur de la moyenne empirique $\overline{X_n}$ est un estimateur sans biais de θ.

2 ▹ Considérons le cas où $X \hookrightarrow \mathscr{U}([0,\theta])$ et $M_n = \max\{X_1,\dots,X_n\}$.

On a déjà vu que M_n est une variable à densité et qu'une densité f_θ est donnée par :

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

Puisque M_n est bornée, l'espérance existe et :

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

donc :

$$b_\theta(M_n) = \mathbb{E}_\theta(M_n) - \theta = -\frac{1}{n+1}\theta$$

donc M_n est un estimateur biaisé de θ.

Cependant, par linéarité de l'espérance, on en déduit que $\frac{n+1}{n}M_n$ est un estimateur sans biais de θ.

Remarque

On dit qu'un estimateur T_n de θ est ***asymptotiquement sans biais*** lorsque :

$$b_n(T_n) \xrightarrow[n\to+\infty]{} 0.$$

Dans l'exemple qui précède, M_n est estimateur asymptotiquement sans biais de θ.

Exercice C-145

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

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

Déterminer $\mathbb{E}(S_n)$, $\mathbb{E}(S_n^2)$ et $\mathbb{E}(T_n)$.

En déduire un estimateur sans biais de la variance de cette loi de Bernoulli.

Exercice C-146

On suppose que X admet une espérance μ et une variance σ^2. On pose :

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

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 XIII-4

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 $g(\theta)$:

$$\forall\theta\in\Theta,\ \forall\varepsilon>0,\ \mathbb{P}_\theta\left(|T_n-g(\theta)|\geqslant\varepsilon\right)\xrightarrow[n\to\infty]{}0.$$

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

Exercice C-147

Considérons X une variable aléatoire dont la loi est uniforme sur $[0,\theta]$, où le paramètre θ est inconnu.

Montrer que $T_n=2\overline{X}_n$ est un estimateur convergent de θ.

Exercice C-148

Soit X une variable à densité donnée pour $\theta>0$ par :

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

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

Vérifier ensuite que cet estimateur est convergent.

Remarque

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 $g(\theta)$ et si f est une fonction continue sur $\mathbb{R}$ alors $(f(T_n))_{n\in\mathbb{N}}$ est une suite d'estimateurs convergente de $f(g(\theta))$.

Exemple

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

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

De même $\ln(X)$ admet un moment d'ordre 2 (donc une variance).

D'après la loi faible des grands nombres, $(\sum_{i=1}^n\ln(X_i))/n$ est un estimateur convergent de $\ln(\theta)-1$.

La fonction $f:t\mapsto\exp(t+1)$ est continue, donc l'estimateur suivant est un estimateur convergent de θ :

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

Proposition XIII-5 (condition suffisante)

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

$$\mathbb{E}(T_n) \xrightarrow[n\to\infty]{} g(\theta) \text{ et } \mathbb{V}(T_n) \xrightarrow[n\to\infty]{} 0.$$

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

Démonstration

Soit $\theta \in \Theta$ et $\varepsilon > 0$. Puisque $(T_n - g(\theta))^2 \geqslant 0$, l'inégalité de Markov donne :

$$\mathbb{P}\left((T_n - g(\theta))^2 \geqslant \varepsilon^2\right) \leqslant \frac{\mathbb{E}\left((T_n - g(\theta))^2\right)}{\varepsilon^2}$$

d'où :

$$\mathbb{P}(|T_n - g(\theta)| \geqslant \varepsilon) \leqslant \frac{\mathbb{E}\left((T_n - g(\theta))^2\right)}{\varepsilon^2}.$$

De plus, on a :

$$\begin{aligned}\mathbb{E}\left((T_n - g(\theta))^2\right) &= \mathbb{E}(T_n^2) - 2g(\theta)\mathbb{E}(T_n) + g(\theta)^2 \\ &= \mathbb{V}(T_n) + \mathbb{E}(T_n)^2 - 2g(\theta)\mathbb{E}(T_n) + g(\theta)^2 \\ &\xrightarrow[n\to+\infty]{} 0 + g(\theta)^2 - 2g(\theta)g(\theta) + g(\theta)^2 = 0\end{aligned}$$

donc, par théorème d'encadrement :

$$\mathbb{P}(|T_n - g(\theta)| \geqslant \varepsilon) \xrightarrow[n\to+\infty]{} 0.$$

Exercice C-149

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 :

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

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.

Exemples

1 ▹ L'estimateur de la moyenne empirique est convergent si X admet une variance.

2 ▹ ***Questions sensibles lors d'un sondage d'opinion –***

Dans un sondage d'opinion, les personnes interrogées peuvent refuser de répondre honnêtement. 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 alors 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 la probabilité t que la personne réponde sans mentir (t est connu, 1/6 dans l'exemple du dé). Pour tout $i \in [\![1,n]\!]$, on note X_i la v.a. de Bernoulli valant 1 le i-ème sondé répond positivement et 0 sinon.

Soit A l'événement « le sondé i ne ment pas» alors la formule de probabilités totales donne :

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

On sait 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 $r = tp + (1-t)(1-p)$. Si $t \neq \frac{1}{2}$, on a :

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

Notons :

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

alors T_n est un estimateur convergent de p. En effet :

- $\overline{X_n}$ est un estimateur convergent de r et f est continue donc $T_n = f\left(\overline{X_n}\right)$ est un estimateur convergent de $f(r) = p$;
- sinon, par linéarité de l'espérance, $\mathbb{E}(T_n) = p$ et T_n est un estimateur sans biais de p, et :

$$\mathbb{V}(T_n) = \frac{r(1-r)}{n(2t-1)^2} \xrightarrow[n\to\infty]{} 0$$

et le résultat découle de la condition suffisante énoncée ci-dessus.

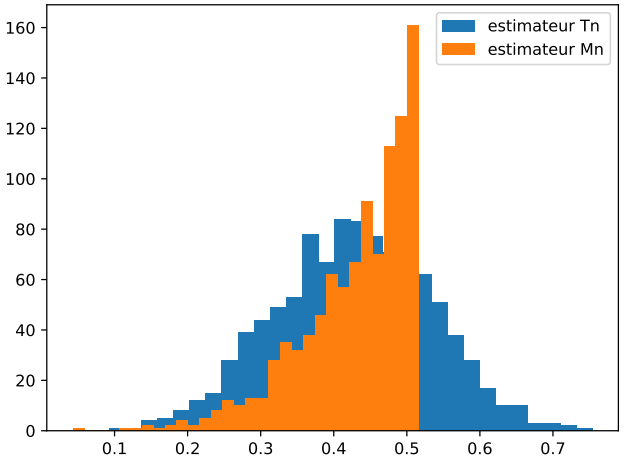
3 ▹ Considérons une v.a. X de loi $\mathscr{U}([0,\theta])$ et les deux estimateurs sans biais de θ vus précédemment :

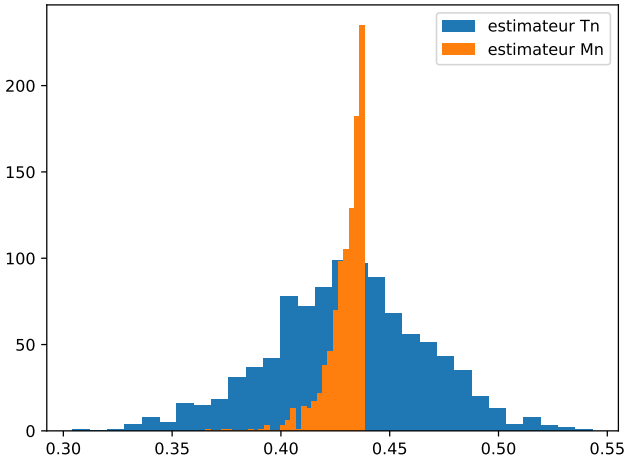
$$\overline{X_n} = \frac{2}{n}\sum_{i=1}^{n} X_i \quad \text{et} \quad \widetilde{M_n} = \frac{n+1}{n}\max(X_1,\dots,X_n).$$

On va effectuer une simulation.

```
import numpy as np
import numpy.random as rd
import matplotlib.pyplot as plt

theta = 0.43 # le paramètre "inconnu"

def estimateurT(n):
    E = [theta*rd.random() for _ in range(n)]
    return 2*np.sum(E)/n

def estimateurM(n):
    t = [theta*rd.random() for _ in range(n)]
    return (n+1)/n * max(t)

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(LT,30,label='estimateur Tn')
    plt.hist(LM,30,label='estimateur Mn')
    plt.legend()
    plt.show()
```

Les instructions `histogrammes(5)` et `histogrammes(50)` donnent les graphiques ci-dessous :





Le meilleur estimateur sans biais est celui qui a la variance la plus faible (celui avec le risque le plus faible que l'échantillon de données soit loin de l'espérance).

B - Estimation par intervalle de confiance

B.1 - 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 $g(\theta)$.

La démarche de l'estimation par intervalle de confiance consiste à trouver un intervalle aléatoire qui contienne $g(\theta)$ avec une probabilité minimale donnée.

Dans tout ce paragraphe, $(U_n)_{n\in\mathbb{N}^*}$ et $(V_n)_{n\in\mathbb{N}^*}$ désigneront deux suites d'estimateurs de $g(\theta)$ telles que, pour tous $\theta \in \Theta$ et $n \in \mathbb{N}^*$, $\mathbb{P}_\theta(U_n \leqslant V_n) = 1$.

Définition XIII-6

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

$$\mathbb{P}_\theta\left(g(\theta) \in [U_n, V_n]\right) \geqslant 1-\alpha.$$

On dit que l'intervalle $[U_n, V_n]$ 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

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

$$u_n = U_n(x_1, x_2, \ldots, x_n) \quad \text{et} \quad v_n = V_n(x_1, x_2, \ldots, x_n).$$

On construit ainsi un intervalle aléatoire $[u_n, v_n]$ dans lequel $g(\theta)$ à une probabilité supérieure à $1-\alpha$ de se trouver.

On considère souvent les cas $\alpha = 0,05$ ou $\alpha = 0,01$ (*i.e.* des niveaux de confiance de 95% ou 99%).

2 ▹ ***Utilisation de l'inégalité de Bienaymé-Tchebychev –***

Soit T_n une variable aléatoire admettant une variance, l'inégalité de Bienaymé-Tchebychev s'écrit :

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

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

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

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

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

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

Exemples

1 ▹ ***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.

Pour tout entier $n \geqslant 1$, on note X_n la v.a. de Bernoulli de paramètre p égale à 1 si la n-ième personne répond positivement et 0 sinon :

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

Soit $\overline{X}_n$ la moyenne empirique alors (du fait de l'indépendance) :

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

L'inégalité de Bienaymé-Tchebychev donne :

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

On vérifie aisément que $p(1-p) \leqslant \frac{1}{4}$. On obtient :

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

Donc :

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

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

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

Par exemple, pour $\alpha = 0,05$ et $n = 100$, on a $\varepsilon \approx 0,22$, on obtient que $\left[\overline{X_n} - 0,22, \overline{X_n} + 0,22\right]$ 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 $\overline{X_n} + 0,22$ (c'est très mauvais !).

2 ▹ ***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 $\mathcal{N}\left(\mu, \sigma^2\right)$. 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, c'est un estimateur sans biais et convergent de μ et le résultat de stabilité des lois normales indépendantes donne :

$$\overline{X}_n \hookrightarrow \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \quad \text{et} \quad Y_n = \sqrt{n} \cdot \frac{\overline{X}_n - \mu}{\sigma} \hookrightarrow \mathcal{N}(0,1).$$

Soit $t_\alpha > 0$ positif avec $\Phi(t_\alpha) = 1 - \frac{\alpha}{2}$, alors :

$$\mathbb{P}\left(-t_\alpha \leqslant Y_n \leqslant t_\alpha\right) = \Phi\left(t_\alpha\right) - \Phi\left(-t_\alpha\right) = 1-\alpha$$

d'où :

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

Donc $\left[\overline{X}_n - t_\alpha \frac{\sigma}{\sqrt{n}}, \overline{X}_n + t_\alpha \frac{\sigma}{\sqrt{n}}\right]$ est un intervalle de confiance de μ avec un niveau de confiance égal à 0,95.

Par exemple pour un risque $\alpha = 0,05$, $1 - \frac{\alpha}{2} = 0,975$ et $t_{0,05} = \Phi^{-1}(0,975) \approx 1,96$.

Pour un risque de $\alpha = 0,01$, $1 - \frac{\alpha}{2} = 0,995$ et $t_{0,01} = \Phi^{-1}(0,995) \approx 2,58$.

B.2 - Intervalle de confiance asymptotique

Définition XIII-7

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

$$\mathbb{P}_\theta\left(g(\theta) \in [U_n, V_n]\right) \geqslant 1-\alpha_n \qquad \text{et} \qquad \alpha_n \xrightarrow[n\to+\infty]{} \alpha.$$

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*** $1-\alpha$**.**

Exemple (intervalle de confiance asymptotique du paramètre d'une loi de Bernoulli)
Soit $(X_1, X_2, \dots, X_n)$ un n-échantillon issu d'une loi de Bernoulli de paramètre p. Par indépendance, on a :

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

et par le théorème central limite, on a :

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

On en déduit si $\Phi(t_\alpha) = 1 - \frac{\alpha}{2}$:

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

L'encadrement $0 \leqslant p(1-p) \leqslant \frac{1}{4}$ permet alors d'écrire :

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

donc (par croissance de la probabilité) :

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

Pour tout $n \in \mathbb{N}$, on pose α_n tel que $1 - \alpha_n$ soit le terme de gauche dans l'inégalité précédente alors :

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

Donc $\left[\overline{X_n} - \frac{t_\alpha}{2\sqrt{n}}, \overline{X_n} + \frac{t_\alpha}{2\sqrt{n}}\right]$ est un intervalle de confiance asymptotique de p au niveau de confiance $1 - \alpha$.

Ce chapitre est essentiellement issu des cours de MM. Vienney (Lycée Champollion) et Fiszka (Lycée Saint Louis).