Chapitre 15

Estimation

Dans les chapitres de probabilité précédents, on savait toujours, dès qu'on avait une variable aléatoire, quelle était sa loi.

En pratique, les statisticiens savent souvent de quel type de loi sera la variable aléatoire étudiée, mais ignorent son ou ses paramètres.

EXEMPLE

On souhaite modéliser la durée de vie T d'un appareil. On sait que T va suivre une loi exponentielle, mais on ignore quel sera le paramètre.

L'objectif de la statistique inférentielle est de déterminer ce paramètre.

15.1 Vocabulaire des statistiques

Définition 15.1.1
On appelle espace des paramètres*, noté Θ, l'ensemble des paramètres (réels ou vectoriels) possibles pour la famille de lois étudiée.*

EXEMPLE

L'espace des paramètres pour la famille de lois $(\mathcal{E}(\theta))_\theta$ est
L'espace des paramètres pour la famille de lois $(\mathcal{N}(\mu, \sigma^2))_{(\mu,\sigma^2)}$ est

Définition 15.1.2
On appelle n-échantillon un vecteur aléatoire $(X_1, \ldots, X_n)$ de variables indépendantes et identiquement distribuées. Leur loi commune est appelée loi-mère *de l'échantillon.*

On appelle réalisation *de cet échantillon toute réalisation du vecteur $(X_1, \ldots, X_n)$,* i.e. *tout vecteur $(X_1(\omega), \ldots, X_n(\omega))$ pour $\omega \in \Omega$.*

Chaque valeur possible de $\theta \in \Theta$ donne naissance à un espace probabilité, qu'on notera $(\Omega_\theta, \mathcal{A}_\theta, P_\theta)$.

De même, on notera E_θ et V_θ l'espérance et la variance pour la probabilité P_θ.

Pour trouver la valeur du paramètre θ, on utilisera des *estimateurs*

Définition 15.1.3
Soit $g : \Theta \to \mathbb{R}$.

On appelle estimateur de $g(\theta)$ d'ordre n *toute variable aléatoire de la forme* , *où* $(X_1, \ldots, X_n)$ *est un* n*-échantillon et* $\varphi : \mathbb{R}^n \to \mathbb{R}$.

NOTA

On estime $g(\theta)$, car dans certains cas, on ne veut pas estimer directement le paramètre, mais une fonction de ce paramètre.

EXEMPLE

Pour la famille de lois $(\mathcal{U}([a,b]))_{a,b}$, on peut vouloir estimer l'étendue de la loi, *i.e.* la valeur de $b - a$.

NOTA

Dans la plupart des cas, c'est quand même le paramètre lui-même qu'on veut estimer, et on choisira donc

Attention ! Un estimateur, malgré son nom, n'est pas toujours un bon moyen pour estimer la valeur souhaitée : le reste de cette partie servira justement à déterminer qui sont les bons estimateurs.

Pour l'instant, un estimateur n'est qu'une fonction de l'échantillon, dont la seule contrainte est de ne pas dépendre de θ.

EXEMPLE

La durée de vie d'un appareil suit une loi exponentielle de paramètre θ inconnu.
Soit $(X_1, \ldots, X_n)$ un échantillon de loi-mère $\mathcal{E}(\theta)$. Des estimateurs de $\frac{1}{\theta}$ sont par exemple

$T_1 =$; $T_2 =$; $T_3 =$

Le premier estimateur de l'exemple a un rôle particulier :

Définition 15.1.4
On appelle moyenne empirique *l'estimateur*

$$\overline{X_n} =$$

Enfin, on pourra aussi définir des suites d'estimateurs :

Définition 15.1.5
On appelle suite d'estimateurs de $g(\theta)$ *la donnée d'un estimateur* T_n *de* $g(\theta)$ *d'ordre* n *pour tout* n*. On aura donc*

$$\forall n \in \mathbb{N}^*,\ T_n = \varphi_n(X_1, \ldots, X_n).$$

15.2 Étude des estimateurs

Dans cette partie, on va voir trois critères qui permettront de déterminer qui sont les "bons" estimateurs.

15.2.1 Biais

Le rôle du biais est d'indiquer si l'estimateur prend des valeurs, en moyenne, autour de la valeur cherchée.

Définition 15.2.1
Soit T_n *un estimateur de* $g(\theta)$*. On suppose que* T_n *admet une espérance pour tout* $\theta \in \Theta$*.*

On appelle biais de T_n en $g(\theta)$ *la valeur*

$$b_\theta(T_n) =$$

Si le biais d'un estimateur est nul, on dira qu'il est sans biais. Dans le cas contraire, on dira qu'il est biaisé.

Proposition 15.2.2
Si la loi-mère d'un échantillon $(X_1, \ldots, X_n)$ *admet une espérance* m_θ*, alors la moyenne empirique*

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

est un estimateur sans biais de m_θ*.*

Démonstration. Il suffit d'utiliser la linéarité de l'espérance pour obtenir $E_\theta(\overline{X_n}) =$. □

EXEMPLE

On voudrait trouver un estimateur sans biais de la variance, dans la cas où la loi-mère possède une espérance m et une variance σ^2 *.

L'intuition nous propose d'utiliser

$$\widetilde{S_n} =$$

Simplifions $\widetilde{S_n}$ pour calculer son espérance :

$\widetilde{S_n} =$

$=$

$=$

$=$

On a donc

$$E_\theta(\widetilde{S_n}) =$$

Finalement, le biais de $\widetilde{S_n}$ est

$$b_\theta(\widetilde{S_n}) =$$

En fait, un estimateur sans biais de la variance est

qui est la formule utilisée par Scilab pour calculer la variance *via* la commande `stdev`.

Définition 15.2.3
On dit qu'une suite d'estimateurs $(T_n)_n$ *est* asymptotiquement sans biais *si*

EXEMPLE

La suite $(\overline{X_n})_n$ est asymptotiquement sans biais ; la suite $(b_\theta(\overline{X_n}))_n$ est constante égale à 0.
La suite $(\widetilde{S_n})_n$ est asymptotiquement sans biais ; on a

$$b_\theta(\widetilde{S_n}) =$$

15.2.2 Risque quadratique

Le rôle du risque quadratique est d'indiquer si les valeurs prises par l'estimateur sont dispersées autour de la valeur cherchée.

Définition 15.2.4
Soit T_n un estimateur de $g(\theta)$. On suppose que T_n admet une variance pour tout $\theta \in \Theta$.

On appelle risque quadratique de T_n en θ *le nombre*

$$r_\theta(T_n) =$$

EXEMPLE

Supposons que la loi-mère de l'échantillon admette une espérance m et une variance σ^2. Alors la moyenne quadratique $\overline{X_n}$ a pour rique quadratique en m

$$\begin{aligned} r_\theta(\overline{X_n}) &= \\ &= \\ &= \end{aligned}$$

Proposition 15.2.5 : Décomposition biais-variance du risque quadratique
Soit T_n un estimateur admettant un risque quadratique. Alors

Démonstration. La variable aléatoire $T_n - g(\theta)$ admet une variance, et par la formule de Huygens

$$V_\theta(T_n - g(\theta)) =$$

Mais $g(\theta)$ est une constante, et donc

$$V_\theta(T_n - g(\theta)) =$$

On a donc

et on retrouve le résultat cherché. □

NOTA

On note que si le risque quadratique est faible, alors le biais aussi.

EXERCICE

On veut estimer le paramètre p d'une loi de Bernoulli. Calculer le risque quadratique pour p de $\overline{X_n}$.
Calculer le risque quadratique pour p de $\frac{1+X_1+\cdots+X_n}{n+2}$.
Quel estimateur a le risque quadratique le plus faible ?

15.2.3 Convergence

Définition 15.2.6
Une suite d'estimateurs $(T_n)_n$ de $g(\theta)$ est dite convergente *si*

Proposition 15.2.7
Supposons que la loi-mère admette une espérance m et une variance σ^2. Alors la moyenne empirique $\overline{X_n}$ est un estimateur sans biais convergent de m.

Démonstration. On a déjà vu que c'était un estimateur sans biais. La
permet de conclure sur la convergence. □

Proposition 15.2.8
Soit (T_n) une suite d'estimateurs de $g(\theta)$ convergente, et soit $f : \mathbb{R} \to \mathbb{R}$ continue.

Alors $(f(T_n))_n)$ est une suite d'estimateurs convergente de

EXEMPLE

Un problème des sondages est l'honnêteté des sondés ; certaines questions génantes (par exemple : consommez-vous du cannabis ?) peuvent les inciter à mentir. Pour pallier à ce problème, on utilise parfois la méthode suivante :

- On pose une question au sondé, qui attend "oui" ou "non" comme réponse.
- Il lance en secret un dé à 6 faces.
- S'il obtient un 6, il répond "oui". Sinon, il dit la vérité.

Notons p la proportion réelle de personnes consommant du cannabis.
La probabilité qu'une personne réponde "oui" au sondage est donc

$$q =$$

Alors la fréquence de "oui" est un estimateur sans biais convergent de q (c'est la moyenne empirique), et donc

$$T_n =$$

est un estimateur convergent de p. De plus, $E(T_n) =$, donc T_n est

Proposition 15.2.9
Soit $(T_n)_n$ une suite d'estimateurs de $g(\theta)$ admettant un risque quadratique. Si

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

Démonstration. $(T_n - g(\theta))^2$ admet une espérance, et donc par l'inégalité de Markov

□

15.3 Intervalles de confiance

Dans toute cette partie, U_n et V_n désigneront deux estimateurs (ou deux suites d'estimateurs) d'ordre n du paramètre $g(\theta)$ cherché, tels que $U_n \leqslant V_n$ presque sûrement.

Définition 15.3.1
Soit $\alpha \in]0,1[$. On dit que $[U_n, V_n]$ est un intervalle de confiance de $g(\theta)$ au niveau de risque

α (ou au niveau de confiance $1-\alpha$) si

EXEMPLE

On veut estimer le paramètre p d'une loi de Bernoulli, avec la moyenne empirique. On souhaite déterminer un intervalle de confiance de p au niveau de risque 0.05.
On a déjà vu que pour une loi de Bernoulli, on peut trouver la majoration suivante dans l'inégalité de Bienaymé-Tchebychev

$$P(|\overline{X_n} - p| \geqslant \varepsilon) \leqslant$$

On en déduit

Pour avoir un niveau de risque de 0.05, on doit donc avoir
Deux choix s'offrent alors :

- En connaissant n, trouver la valeur de ε :

  Si $n = 100$, on trouve alors $\varepsilon \geqslant 0.22$. Un intervalle de confiance de p au niveau de risque 0.05 est donc

  Si $n = 1000$, on trouve

  NOTA

  On note que ces deux intervalles de confiance ont une amplitude beaucoup trop grande pour être utilisables en pratique.

- Trouver la taille de l'échantillon à observer pour assurer un intervalle d'amplitude fixée.

  Pour $\varepsilon = 0.01$, il faut alors $n \geqslant$.

NOTA

On rappelle que l'inégalité de Bienaymé-Tchebychev donne une majoration assez grossière. On utilisera d'autres majoration, qui seront plus fines, mais qui dépendront de la loi choisie.

Définition 15.3.2
Soit $\alpha \in]0,1[$. On dit que $[U_n, V_n]$ est un intervalle de confiance asymptotique de $g(\theta)$ au niveau de confiance $1-\alpha$ si pour tout $\theta \in \Theta$, il existe une suite $(\alpha)_n \in [0,1]^n$ de limite α telle que

Proposition 15.3.3 : Intervalle de confiance pour une loi de Bernoulli
Un intervalle de confiance asymptotique de p au niveau de confiance $1-\alpha$ est

où t_α est l'unique réel vérifiant

Démonstration. Le théorème central limite nous affirme que

.

Donc,

Il nous faut donc $\Phi(b) - \Phi(a) \geqslant 1-\alpha$.

L'usage veut choisir $a = -b$, et donc on cherche b tel que $2\Phi(b) - 1 = 1-\alpha$, *i.e.* $\Phi(b) = 1 - \frac{\alpha}{2}$.

Comme Φ est une bijection, il existe un unique tel b, qu'on note t_α.

On a alors

$\Leftrightarrow$

$\Leftrightarrow$

En notant à nouveau que $p(1-p) \leqslant \frac{1}{4}$, on obtient l'intervalle voulu. □

NOTA

On retiendra quelques valeurs de t_α : $t_{0.05} \simeq$ et $t_{0.01} \simeq$.

On peut en fait faire un raisonnement analogue pour n'importe quelle loi admettant une variance. On se limitera aux lois admettant un moment d'ordre 4.

Proposition 15.3.4
Si la loi-mère des X_n admet un moment d'ordre 4, alors un intervalle de confiance asymptotique de l'espérance au niveau de risque α est donné par

où $\widetilde{S_n}$ est la variance empirique.

Démonstration. Les X_i^2 ont un moment d'ordre 2, et donc d'après la loi faible des grands nombres

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 \xrightarrow{P}$$

Mais on a d'autre part $E(X_1^2) =$.

On a vu que $\overline{X_n}$ est un estimateur convergent de m, et donc $\overline{X_n}^2$ est un estimateur convergent de m^2. Finalement [†], on a

$$\widetilde{S_n} =$$

$\widetilde{S_n}$ est un estimateur (biaisé) convergent de la variance. On a donc

Le théorème central limite nous donne la convergence en loi de $\frac{\sqrt{n}}{\sigma}(\overline{X_n} - m)$ vers une variable $X \hookrightarrow \mathcal{N}(0,1)$, et donc, par le théorème de Slutsky

Soit t_α comme dans la proposition précédente. Alors

puis on obtient le résultat désiré. □

†. Exercice : montrer que la somme ou différence d'estimateurs convergents reste un estimateur convergent

15.4 Exercices

Exercice 1. Soit (X_n) une suite de variables indépendantes de loi $\mathcal{E}(\lambda)$. On pose

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

(i) Calculer le biais de chacun de ces estimateurs de $\frac{1}{\lambda}$.

(ii) Comparer leurs risques quadratiques en tant qu'estimateurs de $\frac{1}{\lambda}$.

Exercice 2. Soient $X_1, \ldots, X_n$ des variables aléatoires indépendantes et identiquement distribuées telles que

$$P(X_1 = -1) = (1-p)^2,\ P(X_1 = 0) = 2p(1-p),\ P(X_1 = 1) = p^2.$$

On cherche dans la suite à estimer $p \in]0,1[$.

Montrer que $Z_n = \displaystyle\sum_{k=1}^{n} \frac{1+X_k}{2n}$ est un estimateur sans biais et convergent de p.

Exercice 3. Soit $a > 0$. On pose

$$f_a : \begin{array}{ccc} \mathbb{R} & \longrightarrow & \mathbb{R} \\ x & \longmapsto & \begin{cases} 0 & \text{si } x < 1 \\ \dfrac{a}{x^{a+1}} & \text{si } x \geqslant 1 \end{cases} \end{array}$$

(i) Montrer que f_a est une densité.

On considère une suite (X_n) de variables indépendantes et identiquement ditribuées de densité f_a, a inconnu.

Pour $n \in \mathbb{N}^*$, on définit une fonction L_n, appellée *vraisemblance*

$$L_n : \begin{array}{ccc} [1,\infty[^n \times \mathbb{R}_+^* & \longrightarrow & \mathbb{R} \\ (x_1, \ldots, x_n, a) & \longmapsto & \prod_{i=1}^n f_a(x_i) \end{array}.$$

b) Pour $x_1, \ldots, x_n$ fixés, montrer que $a \mapsto L_n(x_1, \ldots, x_n, a)$ possède un maximum atteint en un unique réel a que l'on exprimera en fonction des $x_1, \ldots, x_n$.
Indication : On pourra étudier la fonction $a \mapsto \ln L_n(x_1, \ldots, x_n, a)$.

On a donc écrit $a = \varphi_n(x_1, \ldots, x_n)$. On pose alors

$$T_n = \varphi_n(X_1, \ldots, X_n)\ ;$$

T_n s'appelle *estimateur du maximum de vraisemblance*.

c) Montrer que pour tout k, $\ln(X_k)$ suit une loi exponentielle dont on précisera le paramètre. En déduire une densité de $S_n = \sum_{i=1}^n \ln(X_i)$.

d) Exprimer T_n en fonction de S_n. En déduire $E(T_n)$ et $V(T_n)$.

e) Montrer que T_n est un estimateur asymptotiquement sans biais et convergent de a.

Exercice 4. Soit X une variable aléatoire dont la loi dépend d'un paramètre θ.

Soient T_1, T_2 deux estimateurs indépendants sans biais de θ, de variances respectives V_1 et V_2. Pour tout réel a, on pose

$$\Theta_a = aT_1 + (1-a)T_2.$$

Θ_a est-il un estimateur sans biais de θ ?

Déterminer a pour que la variance de Θ_a soit minimale. Quelle est la valeur de cette variance ?

Exercice 5. Soient $a > 0$ et $X \hookrightarrow \mathcal{U}([0, 2a])$.

(i) Soit $n \in \mathbb{N}^*$. On considère n variables indépendantes de même loi que X. On pose $M_n = \max(X_1, \ldots, X_n)$.
Déterminer la loi de M_n et calculer son espérance et sa variance.

(ii) En déduire que $U_n = \frac{n+1}{2n} M_n$ est un estimateur sans biais de $E(X)$. Est-il préférable à l'estimateur $\overline{X_n}$?

Exercice 6. On suppose que la probabilité qu'un individu contagieux transmette un virus à un individu sain est $p \in]0, 1[$ inconnu, et que l'on cherche à évaluer.

Soit (Y_n) une suite de variables alatoires indépendantes et identiquement distribuées suivant une loi de Bernoulli de paramètre p.

(i) On pose

$$\overline{Y_n} = \frac{1}{n} \sum_{i=1}^{n} Y_i.$$

Montrer que $\overline{Y_n}$ est un estimateur sans biais de p, et déterminer son risque quadratique.

(ii) À l'aide de l'inégalité de Bienaymé-Tchebychev, montrer que $\left[\overline{Y_n} - \sqrt{\frac{5}{n}}, \overline{Y_n} + \sqrt{\frac{5}{n}}\right]$ est un intervalle de confiance de p au niveau de confiance 0.95.

Exercice 7. Soit (X_n) une suite de variables indépendantes et identiquement distribuées d'espérance μ inconnue et de variance σ^2 connue.

Déterminer, à l'aide du théorème central limite, un intervalle de confiance asymptotique de μ au niveau de confiance $1 - \alpha$.

Exercice 8. Soient $n \geqslant 1$ et $(X_1, \ldots, X_n)$ un échantillon indépendant et identiquement distribué de la loi de Poisson de paramètre $\lambda > 0$ inconnu. On cherche à estimer λ par un intervalle de confiance. On pose

$$\overline{X_n} = \frac{1}{n} \sum_{i=1}^{n} X_i \text{ et } T_n = \sqrt{n} \frac{\overline{X_n} - \lambda}{\sqrt{\lambda}}.$$

À l'aide de T_n, déterminer, pour n grand, un intervalle de confiance de λ au risque α donné.