

L3 Economie

Statistiques inférentielles

Chapitre 4 : Intervalles de confiance

Cécile Durot
cecile.durot@gmail.com

Université Paris Nanterre

Cadre d'étude

Dans toute la suite, nous supposons que les données collectées sont des nombres réels $x_1, \dots, x_n$, et que le procédé de collecte est suffisamment proche du tirage aléatoire simple pour que l'on puisse supposer que :

Les données $x_1, \dots, x_n$ sont les réalisations de v.a.r. $X_1, \dots, X_n$ indépendantes et de même loi (i.i.d.).

On notera $\theta_0 \in \mathbb{R}$ le paramètre d'intérêt, X une variable aléatoire abstraite de même loi que $X_1, \dots, X_n$ et P_X cette loi.

Introduction

Construction d'intervalles de confiance

Construction d'intervalles de confiance asymptotiques

En utilisant le TLC

Dans le cadre des familles exponentielles

Problématique

Un estimateur de θ_0 fournit une valeur approchée pour θ_0 mais l'erreur commise n'est alors pas quantifiée.

Définition : Soit $\alpha \in]0, 1[$ fixé (typiquement, $\alpha = 0.05$).

Un intervalle de confiance de niveau de confiance $1 - \alpha$ (ce qu'on notera $IC(\alpha)$) est un intervalle $[A_n, B_n]$ où A_n et B_n sont des variables aléatoires s'exprimant comme fonction de $X_1, \dots, X_n$ uniquement et telles que

$$P(\theta_0 \in [A_n, B_n]) \geq 1 - \alpha.$$

Deux erreurs quantifiées :

- On ne donne pas la valeur exacte de θ_0 mais un intervalle,
- θ_0 n'appartient à l'intervalle qu'avec une probabilité $1 - \alpha$.

Si l'on cherche à diminuer l'une, on augmente l'autre. C'est pourquoi il faut choisir un intervalle dont la probabilité associée est aussi proche que possible de $1 - \alpha$ (si possible, égale à $1 - \alpha$).

Cadre asymptotique

Dans certains cas, il n'est pas possible (ou difficile) de construire un $IC(\alpha)$. Dans ce cas, on en cherche une version asymptotique, c'est-à-dire qui est valable lorsque $n \rightarrow \infty$.

Définition : Soit $\alpha \in]0, 1[$ fixé (typiquement, $\alpha = 0.05$).

Un intervalle de confiance asymptotique de niveau de confiance $1 - \alpha$ (ce qu'on notera $ICA(\alpha)$) est un intervalle $[A_n, B_n]$ où A_n et B_n sont des variables aléatoires s'exprimant comme fonction de $X_1, \dots, X_n$ uniquement, et telles que

$$\lim_{n \rightarrow \infty} P(\theta_0 \in [A_n, B_n]) \geq 1 - \alpha.$$

Introduction

Construction d'intervalles de confiance

Construction d'intervalles de confiance asymptotiques

En utilisant le TLC

Dans le cadre des familles exponentielles

Méthode de construction d'un $IC(\alpha)$

Si possible, on utilisera la méthode suivante de construction :

- ❶ Construire un bon estimateur $\hat{\theta}_n$ de θ_0 .
- ❷ Déterminer la loi de $\hat{\theta}_n$.
- ❸ Construire à partir de $\hat{\theta}_n$ une variable aléatoire T_n dont la loi est entièrement connue. La définition de T_n implique nécessairement θ_0 mais ne doit impliquer aucun autre paramètre inconnu.
- ❹ Donner un intervalle I_n auquel appartient T_n avec une probabilité d'au moins $1 - \alpha$.
- ❺ "Inverser" l'intervalle, c'est-à-dire déterminer un intervalle $[A_n, B_n]$ tel qu'on ait l'équivalence $T_n \in I_n \iff \theta_0 \in [A_n, B_n]$.

Alors, $[A_n, B_n]$ est un $IC(\alpha)$ pour θ_0 .

Exemple espérance Gaussienne - variance connue

Supposons que $X_1, \dots, X_n$ sont i.i.d. Gaussiennes de variance connue σ^2 et que le paramètre d'intérêt est $\theta_0 = E(X)$.

- ❶ L'estimateur empirique (et EMV) de θ_0 est $\bar{X}_n = \sum_{i=1}^n X_i/n$.
- ❷ $\bar{X}_n \sim \mathcal{N}(\theta_0, \sigma^2/n)$.
- ❸ $T_n = \sqrt{n} \frac{\bar{X}_n - \theta_0}{\sigma} \sim \mathcal{N}(0, 1)$.
- ❹ Notons q_α le quantile d'ordre $1 - \alpha/2$ de la loi $\mathcal{N}(0, 1)$. Alors,

$$P\left(\sqrt{n} \frac{\bar{X}_n - \theta_0}{\sigma} \in [-q_\alpha, q_\alpha]\right) = 1 - \alpha.$$

- ❺ On a $\sqrt{n} \frac{\bar{X}_n - \theta_0}{\sigma} \in [-q_\alpha, q_\alpha]$ si et seulement si

$$\theta_0 \in \left[\bar{X}_n - q_\alpha \frac{\sigma}{\sqrt{n}}; \bar{X}_n + q_\alpha \frac{\sigma}{\sqrt{n}}\right]$$

donc ce dernier intervalle est un $IC(\alpha)$ pour θ_0 .

Exemple espérance Gaussienne - variance inconnue

Supposons que $X_1, \dots, X_n$ sont i.i.d. Gaussiennes de variance inconnue σ^2 et que le paramètre d'intérêt est $\theta_0 = E(X)$. On ne peut pas considérer

$$T_n = \sqrt{n} \frac{\bar{X}_n - \theta_0}{\sigma} \sim \mathcal{N}(0, 1)$$

qui implique σ , à présent inconnu. A la place, on considère

$$\tilde{T}_n = \sqrt{n} \frac{\bar{X}_n - \theta_0}{S_n}$$

où S_n^2 est l'estimateur empirique corrigé de σ^2 . Attention, $\tilde{T}_n$ ne suit pas la loi Gaussienne standard.

Théorème : $\tilde{T}_n \sim \mathcal{T}(n-1)$, la loi de Student à $n-1$ degrés de liberté.

Propriétés de la loi de Student :

- La loi $\mathcal{T}(n-1)$ est une loi possédant une densité paire dont la représentation graphique est en forme de cloche.
- Une v.a.r. de loi $\mathcal{T}(n-1)$ est centrée.
- Il existe une écriture analytique pour la densité de la loi $\mathcal{T}(n-1)$ mais nous ne l'utiliserons pas dans ce cours. En revanche, il n'existe pas d'écriture analytique pour la fonction de répartition. Comme pour la loi Gaussienne ou χ^2 , on a recours à un logiciel ou à une table de valeurs numériques.
- Remarquons que $\tilde{T}_n$ défini précédemment converge en loi vers $\mathcal{N}(0, 1)$ quand $n \rightarrow \infty$, ce qui montre que la loi de Student 'ressemble' à la loi Gaussienne standard quand n est grand.

Si $X_1, \dots, X_n$ sont i.i.d. Gaussiennes de variance inconnue σ^2 de paramètre d'intérêt $\theta_0 = E(X)$, en notant $\tilde{q}_{n,\alpha}$ le quantile d'ordre $1 - \alpha/2$ de la loi $\mathcal{T}(n-1)$, on a

$$P\left(\sqrt{n}\frac{\bar{X}_n - \theta_0}{S_n} \in [-\tilde{q}_{n,\alpha}, \tilde{q}_{n,\alpha}]\right) = 1 - \alpha$$

d'où l'on déduit que

$$\left[\bar{X}_n - \tilde{q}_{n,\alpha} \frac{S_n}{\sqrt{n}}; \bar{X}_n + \tilde{q}_{n,\alpha} \frac{S_n}{\sqrt{n}}\right]$$

est un $IC(\alpha)$ pour θ_0 .

Exemple variance Gaussienne - espérance inconnue

Supposons que $X_1, \dots, X_n$ sont i.i.d. Gaussiennes $\mathcal{N}(m, \sigma^2)$ où m et σ^2 sont inconnues, et que le paramètre d'intérêt est $\theta_0 = \sigma^2$. Soit

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2.$$

Théorème : $\frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2(n-1).$

Avec q_- et q_+ les quantiles d'ordre $\alpha/2$ et $1 - \alpha/2$ de la loi $\chi^2(n-1)$,

$$P\left(\frac{(n-1)S_n^2}{\sigma^2} \in [q_-, q_+]\right) = 1 - \alpha$$

donc :

$$\left[\frac{(n-1)S_n^2}{q_+}, \frac{(n-1)S_n^2}{q_-}\right] \text{ est un } IC(\alpha) \text{ pour } \sigma^2.$$

Introduction

Construction d'intervalles de confiance

Construction d'intervalles de confiance asymptotiques

En utilisant le TLC

Dans le cadre des familles exponentielles

ICA pour l'espérance - variance connue

Supposons que $X_1, \dots, X_n$ sont i.i.d. et possèdent une espérance inconnue θ_0 et une variance $\sigma^2 > 0$ connue. D'après le TLC,

$$\sqrt{n} \frac{\bar{X}_n - \theta_0}{\sigma} \xrightarrow[n \rightarrow \infty]{\mathcal{L}} \mathcal{N}(0, 1).$$

En notant q_α le quantile d'ordre $1 - \alpha/2$ de la loi $\mathcal{N}(0, 1)$, il découle de la définition de la convergence en loi que

$$\lim_{n \rightarrow \infty} P \left(\sqrt{n} \frac{\bar{X}_n - \theta_0}{\sigma} \in [-q_\alpha, q_\alpha] \right) = 1 - \alpha.$$

On a $\sqrt{n} \frac{\bar{X}_n - \theta_0}{\sigma} \in [-q_\alpha, q_\alpha]$ si et seulement si

$$\theta_0 \in \left[\bar{X}_n - q_\alpha \frac{\sigma}{\sqrt{n}}; \bar{X}_n + q_\alpha \frac{\sigma}{\sqrt{n}} \right]$$

donc ce dernier intervalle est un $ICA(\alpha)$ pour θ_0 .

ICA pour l'espérance - variance inconnue

Supposons que $X_1, \dots, X_n$ sont i.i.d. et possèdent une espérance inconnue θ_0 et une variance $\sigma^2 > 0$ inconnue. D'après le TLC combiné au théorème de Slutsky,

$$\sqrt{n} \frac{\bar{X}_n - \theta_0}{S_n} \xrightarrow[n \rightarrow \infty]{\mathcal{L}} \mathcal{N}(0, 1).$$

En notant q_α le quantile d'ordre $1 - \alpha/2$ de la loi $\mathcal{N}(0, 1)$, il découle de la définition de la convergence en loi que

$$\lim_{n \rightarrow \infty} P \left(\sqrt{n} \frac{\bar{X}_n - \theta_0}{S_n} \in [-q_\alpha, q_\alpha] \right) = 1 - \alpha.$$

Donc

$$\left[\bar{X}_n - q_\alpha \frac{S_n}{\sqrt{n}}; \bar{X}_n + q_\alpha \frac{S_n}{\sqrt{n}} \right]$$

est un $ICA(\alpha)$ pour θ_0 .

Introduction

Construction d'intervalles de confiance

Construction d'intervalles de confiance asymptotiques

En utilisant le TLC

Dans le cadre des familles exponentielles

Cadre

Supposons qu'on dispose d'une famille de lois indexée par $\theta \in \mathbb{R}$, $\{P_\theta, \theta \in \Theta\}$, à laquelle appartient la loi de X : il existe donc θ_0 tel que $P_X = P_{\theta_0}$. On suppose dans la suite θ_0 identifiable au sens où $P_\theta \neq P_{\theta_0}$ dès lors que $\theta \neq \theta_0$.

On suppose de plus que

- la famille de lois $\{P_\theta, \theta \in \Theta\}$ est exponentielle,
- l'EMV $\hat{\theta}_n$ de θ_0 existe,
- La fonction I d'information de Fisher est continue et vérifie $I(\theta_0) > 0$.

Construction d'un $ICA(\alpha)$ pour θ_0

On sait que dans le cadre précédent,

$$\sqrt{nl(\hat{\theta}_n)}(\hat{\theta}_n - \theta_0) \xrightarrow[n \rightarrow \infty]{\mathcal{L}} \mathcal{N}(0, 1).$$

En notant q_α le quantile d'ordre $1 - \alpha/2$ de la loi $\mathcal{N}(0, 1)$, il découle de la définition de la convergence en loi que

$$\lim_{n \rightarrow \infty} P\left(\sqrt{nl(\hat{\theta}_n)}(\hat{\theta}_n - \theta_0) \in [-q_\alpha, q_\alpha]\right) = 1 - \alpha.$$

On a $\sqrt{nl(\hat{\theta}_n)}(\hat{\theta}_n - \theta_0) \in [-q_\alpha, q_\alpha]$ si et seulement si

$$\theta_0 \in \left[\hat{\theta}_n - \frac{q_\alpha}{\sqrt{nl(\hat{\theta}_n)}}; \hat{\theta}_n + \frac{q_\alpha}{\sqrt{nl(\hat{\theta}_n)}} \right]$$

donc ce dernier intervalle est un $ICA(\alpha)$ pour θ_0 .

Exemples

On note q_α le quantile d'ordre $1 - \alpha/2$ de la loi $\mathcal{N}(0, 1)$.

- Si $P_\theta = \mathcal{B}(\theta)$, $\theta \in]0, 1[$, alors $\hat{\theta}_n = \bar{X}_n$ et $I(\theta) = 1/(\theta(1 - \theta))$.
On obtient l' $ICA(\alpha)$ pour θ_0 :

$$\left[\bar{X}_n - q_\alpha \sqrt{\frac{\bar{X}_n(1 - \bar{X}_n)}{n}}; \bar{X}_n + q_\alpha \sqrt{\frac{\bar{X}_n(1 - \bar{X}_n)}{n}} \right]$$

- Si $P_\theta = \mathcal{P}(\theta)$, $\theta > 0$, alors $\hat{\theta}_n = \bar{X}_n$ et $I(\theta) = 1/\theta$. On obtient l' $ICA(\alpha)$ pour θ_0 :

$$\left[\bar{X}_n - q_\alpha \sqrt{\frac{\bar{X}_n}{n}}; \bar{X}_n + q_\alpha \sqrt{\frac{\bar{X}_n}{n}} \right]$$

- Si $P_\theta = \mathcal{E}(\theta)$, $\theta \in]0, 1[$, alors $\hat{\theta}_n = 1/\bar{X}_n$ et $I(\theta) = 1/\theta^2$. On obtient l' $ICA(\alpha)$ pour θ_0 :

$$\left[\frac{1}{\bar{X}_n} - \frac{q_\alpha \bar{X}_n}{\sqrt{n}}; \frac{1}{\bar{X}_n} + \frac{q_\alpha \bar{X}_n}{\sqrt{n}} \right]$$