

MS BGD: MDI720

Intervalles de confiance

François Portier, Anne Sabourin
Telecom ParisTech

Septembre 2018

1. Intervalle de confiance

Définition

Théorèmes limites

IC pour le modèle linéaire

1. Intervalle de confiance

Définition

Théorèmes limites

IC pour le modèle linéaire

Intervalle de confiance

- Contexte : on a une estimation $\hat{g}(y_1, \dots, y_n)$ d'une grandeur g . On veut un intervalle \hat{I} autour de \hat{g} qui contient g avec une grande probabilité.
- On construit $\hat{I} = [\underline{C}, \overline{C}]$ à partir des observations (y_1, \dots, y_n) : l'intervalle est une variable aléatoire

$$\mathbb{P}(\hat{I} \text{ contient } g) = \mathbb{P}(\underline{C} \leq g \text{ et } \overline{C} \geq g) = 95\%$$

Intervalle de confiance de niveau α

Intervalle de confiance

Un intervalle de confiance de niveau α pour la grandeur g est une fonction de l'échantillon

$$\hat{I} : (y_1, \dots, y_n) \mapsto \hat{I} = [\underline{C}(y_1, \dots, y_n), \overline{C}(y_1, \dots, y_n)]$$

telle que

$$\mathbb{P} \left[g \in \hat{I}(y_1, \dots, y_n) \right] \geq 1 - \alpha$$

Rem: choix classiques $\alpha = 5\%, 1\%, 0.1\%$, etc. Résultant souvent d'un arbitrage complexité des données / nombre d'échantillons

Rem: Dans la suite on notera IC pour Intervalle de Confiance

Exemple : sondage

- Sondage d'une élection à deux candidats : A et B . Le choix du i^{e} sondé suit une loi de Bernoulli de paramètre p , avec $y_i = 1$ s'il vote A , 0 sinon.
- But : estimer p .
- échantillon de taille n : un estimateur raisonnable est alors

$$\hat{p} = \frac{1}{n} \sum_{i=1}^n y_i = \bar{y}_n$$

intervalle de confiance pour p ?

Sondage : intervalle de confiance

- Chercher un intervalle $\hat{I} = [\hat{p} - \delta, \hat{p} + \delta]$ tel que $\mathbb{P}(p \in \hat{I}) \geq 0.95 \Leftrightarrow$ chercher δ tel que $\mathbb{P}[|\hat{p} - p| > \delta] \leq 0.05$
- Ingrédient : inégalité de **Tchebyshev**

$$\boxed{\forall \delta > 0, \quad \mathbb{P}(|X - \mathbb{E}(X)| > \delta) \leq \frac{\text{Var}(X)}{\delta^2}}$$

Pour $X = \hat{p} = \frac{1}{n} \sum_{i=1}^n Y_i$ on a $\mathbb{E}(\hat{p}) = p$ et $\text{Var}(\hat{p}) = \frac{p(1-p)}{n}$:

$$\forall p \in (0, 1), \forall \delta > 0, \quad \mathbb{P}(|\hat{p} - p| > \delta) \leq \frac{p(1-p)}{n\delta^2} \leq \frac{1}{4n\delta^2}$$

Application numérique : pour un IC à 95%, choisir δ tel que $\frac{1}{4n\delta^2} = 0.05$, *i.e.* $\delta = (0.2n)^{-1/2}$. Si $n = 1000$, $\hat{p} = 55\%$:

$$\delta = 0.07 ; \quad \hat{I} = [0.48, 0.62]$$

1. Intervalle de confiance

Définition

Théorèmes limites

IC pour le modèle linéaire

Théorème central limite

- y_1, y_2, \dots , des variables aléatoires *i.i.d.* de carré intégrable.
- μ et σ leur espérance et écart-type théoriques.

Théorème central limite (TCL)

La loi de la moyenne empirique re-normalisée $\sqrt{n} \left(\frac{\bar{y}_n - \mu}{\sigma} \right)$ converge vers une loi normale centrée réduite $\mathcal{N}(0, 1)$

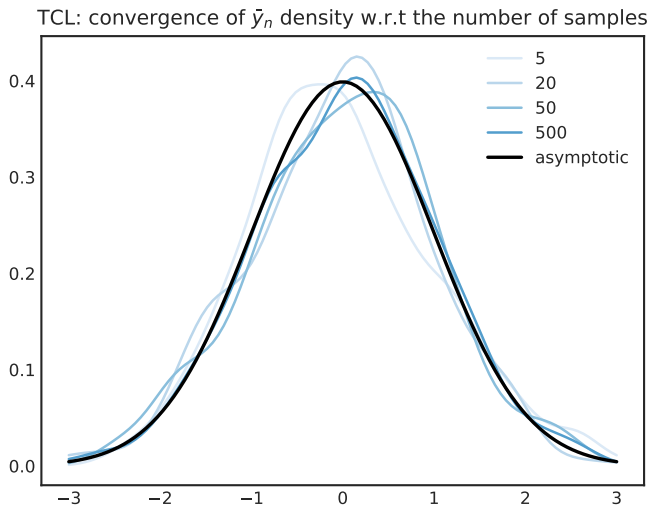
- σ est connu

Lemme de Slutsky

La loi de la moyenne empirique “studentisée” $\sqrt{n} \left(\frac{\bar{y}_n - \mu}{\hat{\sigma}} \right)$ converge vers une loi normale centrée réduite $\mathcal{N}(0, 1)$ quand $\hat{\sigma} \rightarrow \sigma$

Reformulation : $\bar{y}_n \simeq \mathcal{N}(\mu, \hat{\sigma}^2/n)$

Illustration



Intervalles de confiance asymptotiques

- Exemple du sondage : $y_i \in \{0, 1\}$, $n = 1000$,

$$\hat{p} = n^{-1} \sum_{i=1}^n y_i = 0.55$$

- On suppose que n est suffisamment grand pour que

$$\sqrt{n} \left(\frac{\hat{p} - p}{\hat{\sigma}} \right) \sim \mathcal{N}(0, 1)$$

$$\hat{\sigma}^2 = n^{-1} \sum_{i=1}^n (y_i - \hat{p})^2 = \hat{p} - \hat{p}^2$$

- On connaît les quantiles de la loi normale (numériquement)
 $q(1 - 0.05/2) \simeq 1.96$
- D'après le TCL, et l'approximation des quantiles gaussiens

$$\mathbb{P} \left[-1.96 < \sqrt{n} \frac{0.55 - p}{\hat{\sigma}} < 1.96 \right] \approx 0.95$$

nouvel IC : $\hat{I} = [0.52, 0.58]$: meilleur ! (**plus optimiste**)

En Python

Génération des données

```
import numpy as np
from scipy.stats import norm

n = 1000
x = np.random.binomial(1, .5, n)
```

Calcul de l'IC

```
pchap = np.mean(x)
sig = np.sqrt(pchap * (1 - pchap))
alpha = .05
q = norm.ppf(1 - alpha/2)
borneinf = pchap - sig * q / np.sqrt(n)
bornesup = pchap + sig * (1 - q) / np.sqrt(n)
print('IC = [' + str(borneinf) +
      ', ' + str(bornesup) + ' ]')
```

1. Intervalle de confiance

Définition

Théorèmes limites

IC pour le modèle linéaire

IC pour les moindres carrés (I)

Rappel : prenons $X \in \mathbb{R}^{n \times p}$, alors $\hat{\sigma}^2 = \|\mathbf{y} - X\hat{\boldsymbol{\theta}}\|_2^2 / (n - \text{rang}(X))$, estimateur sans biais de la variance. De plus (cf. Poly) :

$$\text{Si } \boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 \text{Id}_n), \text{ alors } T_j = \frac{\hat{\theta}_j - \theta_j^*}{\hat{\sigma} \sqrt{[(X^\top X)^{-1}]_{j,j}}} \sim \mathcal{T}_{n-\text{rang}(X)}$$

où $\mathcal{T}_{n-\text{rang}(X)}$ est une loi dite de Student (de degré $n - \text{rang}(X)$).

Sa densité, ses quantiles, etc. sont calculables numériquement.

IC pour les moindres carrés (II)

Sous l'hypothèse gaussienne, comme

$$T_j = \frac{\hat{\theta}_j - \theta_j^*}{\hat{\sigma} \sqrt{(X^\top X)^{-1}_{jj}}} \sim \mathcal{T}_{n-\text{rang}(X)}$$

et en notant $t_{1-\alpha/2}$ un quantile d'ordre $1 - \alpha/2$ de la loi $\mathcal{T}_{n-\text{rang}(X)}$, alors l'intervalle de confiance suivant est de niveau α

$$\left[\hat{\theta}_j - t_{1-\alpha/2} \hat{\sigma} \sqrt{(X^\top X)^{-1}_{jj}}, \hat{\theta}_j + t_{1-\alpha/2} \hat{\sigma} \sqrt{(X^\top X)^{-1}_{jj}} \right]$$

pour la quantité θ_j^* .

Rem: $\mathbb{P}(|T_j| < t_{1-\alpha/2}) = 1 - \alpha$ car la loi de Student est symétrique

Limites des IC précédents

Dans la partie précédente, l'intégralité des raisonnements repose sur le modèle gaussien ou **l'approximation asymptotique**.

Attention : si le modèle est (trop) faux ou l'échantillon trop petit alors les IC obtenus ne seront pas forcément pertinents.

Alternative possible : *bootstrap*, une méthode non-paramétrique reposant sur le ré-échantillonnage, bien fondée (théoriquement) pour des statistiques régulières telle que la moyenne, les quantiles, etc., (mais pas pour le max ou le min !)

Pour aller plus loin : [Efron et Tibshirani \(1994\)](#)

Références I

- [ET94] B. Efron and R. Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.