Cours de boostrap et ré-échantillonnage

March 30, 2020

Plan

Introduction

But de l'inférence statistique

On a

- $\mathcal{X}_n = (X_1, \ldots, X_n)$ un échantillon i.i.d. de fonction de répartition F
- $\theta(F)$ une quantité d'intérêt, qui dépend de F
- $T(\mathcal{X}_n)$ une statistique, estimateur de $\theta(F)$,

on voudrait connaître

- le biais : $\mathbb{E}_F(T(\mathcal{X}_n)) - \theta(F)$
- la variance : $\mathbb{E}_F(T^2(\mathcal{X}_n)) - \mathbb{E}_F^2(T(\mathcal{X}_n))$
- le MSE (EQM) : $\mathbb{E}_F\Big((T(\mathcal{X}_n) - \theta(F))^2\Big)$
- la loi de $T(\mathcal{X}_n)$: $G^n(x) = \mathbb{P}_F(T(\mathcal{X}_n) \leq x),\ \forall x.$
- etc

pour comparer des estimateurs, connaître leurs qualités, construire des intervalles de confiance...
Problème : toutes ses quantités dépendent de la loi F inconnue !

Ce que l'on sait

On a à disposition la fonction de répartition empirique des X_i.

Fonction de répartition empirique

Pour $\mathcal{X}_n = (X_1, \ldots, X_n)$, la fonction de répartition empirique est définie par

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{X_i \leq x}, \ \forall x.$$

On va considérer les estimateurs par plug-in.

Principe de plug-in

Pour tout paramètre $\theta(F)$ et tout échantillon $\mathcal{X}_n = (X_1, \ldots, X_n)$, on considère l'estimateur par plug-in

$$T(\mathcal{X}_n) = \theta(F_n) = \hat{\theta}$$

$\rightarrow$ exemples : espérance, variance, médiane

Détails

Bootstrap

Bootstrap d'Efron 1982; Efron 1992

Conditionnellement à $\mathcal{X}_n = (X_1, \ldots, X_n)$, on construit des échantillons

$$\mathcal{X}_1^* = (X_{1,1}^* = X_{m_1}, \ldots, X_{1,n}^* = X_{m_n})$$
$$\ldots$$
$$\mathcal{X}_b^* = (X_{b,1}^* = X_{m_{(b-1)n+1}}, \ldots, X_{b,n}^* = X_{m_{bn}})$$
$$\ldots$$

où les m_k ont été tirés aléatoirement et avec remise dans $\{1, \ldots, n\}$.

Loi des $X_{b,j}^*$ conditionnelle à $\mathcal{X}_n$

Conditionnellement $\mathcal{X}_n$, $X_{b,j}^*$ est une v.a. de fonction de répartition F_n, fonction de répartition empirique des $X_1, \ldots, X_n$.

Détails

Estimateurs du bootstrap classique

Soit un paramètre inconnu $\theta(F)$

- Monde réel
  - avec l'échantillon initial $\mathcal{X}_n$, on définit l'estimateur $\hat{\theta} = \theta(F_n) = T(\mathcal{X}_n)$
  - on note G^n la f.d.r. inconnue de $\hat{\theta}$, qui dépend de F (et de n !), inconnue
- Monde bootstrap
  - pour chaque échantillon bootstrapé $\mathcal{X}_b^*$, on définit l'estimateur $\hat{\theta}_b^* = T(\mathcal{X}_b^*)$
  - conditionnellement à F_n, de loi $G^{n,*}$ qui dépend de F_n
  - on estime $G^{n,*}$ par

$$\widehat{G}_{n,B}^*(t) = \frac{1}{B} \sum_{b=1}^{B} \mathbb{1}_{\hat{\theta}_b^* \leq t}$$

Détails

Exemples d'estimateurs bootstrap (1)

Estimation de la loi de $\hat{\theta}$

La f.d.r. G^n de $\hat{\theta} = T(\mathcal{X}_n)$ est définie pour $t \in \mathbb{R}$ par

$$G^n(t) = \int \mathbb{1}_{x \leq t} dG^n(x)$$

elle est estimée par (1ere approximation du bootstrap)

$$G^{n,*}(t) = \int \mathbb{1}_{x \leq t} dG^{n,*}(x) = \mathbb{P}_{F_n}(\hat{\theta}_b^* \leq t)$$

puis (2ieme approximation du bootstrap)

$$\widehat{G}_{n,B}^*(t) = \int \mathbb{1}_{x \leq t} d\widehat{G}_{n,B}^*(x).$$

Exemples d'estimateurs bootstrap (2)

Estimation du biais de $\hat{\theta}$

Le biais de $\hat{\theta}$ est défini par

$$\mathbb{E}_F(T(\mathcal{X}_n)) - \theta(F) = \int x dG^n(x) - \theta(F)$$

est estimé (1ere approximation) par

$$\int x dG^{n,*}(x) - \theta(F_n)$$

puis (2ieme approximation) par

$$\int x d\widehat{G}^*_{n,B}(x) - \theta(F_n) = \frac{1}{B}\sum_{b=1}^{B} \hat{\theta}^*_b - \theta(F_n).$$

Exemples d'estimateurs bootstrap (3)

Estimation de la variance de $\hat{\theta}$

Le biais de $\hat{\theta}$ est défini par

$$\mathbb{E}_F((T(\mathcal{X}_n) - \mathbb{E}_F(T(\mathcal{X}_n))^2) = \int (x - \int x dG^n)^2 dG^n(x)$$

est estimé (1ere approximation) par

$$\int (x - \int x dG^{n,*})^2 dG^{n,*}(x)$$

puis (2ieme approximation) par

$$\int (x - \int x d\widehat{G}^*_{n,B})^2 d\widehat{G}^*_{n,B}(x) = \frac{1}{B}\sum_{b=1}^{B}(\hat{\theta}^*_b - \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^*_b)^2.$$

etc

Eléments de validation asymptotique du bootstrap (1)

Le bootstrap fait deux approximations

$$G^n \overset{(1)}{\to} G^{n,*} \overset{(2)}{\to} \widehat{G}^*_{n,B}.$$

Pour contrôler la deuxième approximation, on utilise une borne de Dvoretsky-Kiefer-Wolfowitz.

Borne DKW, DKW (1956) - Massart (1990)

Si $Z_1, \ldots, Z_N$ est un échantillon i.i.d. de f.d.r. H et H_N est la f.d.r. empirique associée alors

$$\mathbb{P}\left(\sqrt{N}\sup_{x\in\mathbb{R}}|H_N(x) - H(x)| > \epsilon\right) \leq 2\exp(-2\epsilon^2).$$

Application pour le choix de B :
Si on veut que $\sup_{x\in\mathbb{R}}|\widehat{G}^*_{n,B}(x) - G^{n,*}(x)| \leq 0.02$ avec une probabilité plus grande que 0.05, comment choisir B ?

Eléments de validation asymptotique du bootstrap (2)

La première approximation est contrôlée par les développements d'Edgeworth (voir Hall 2013). Si $\hat{\theta}$ est asymptotiquement normal :

$$S = \sqrt{n}\frac{\hat{\theta} - \theta}{\sigma(F)} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$$

avec quelques conditions supplémentaires, on peut montrer que G^n admet un développement d'Edgeworth

$$\mathbb{P}(S \leq x) = G^n(\theta + \sigma x/\sqrt{n}) = \Phi(x) + \frac{1}{n^{1/2}} p(x)\phi(x) + \mathcal{O}(\frac{1}{n}).$$

Dans le monde bootstrap, on peut montrer que si

$$S^* = \sqrt{n}\frac{\hat{\theta}^* - \hat{\theta}}{\sigma(F)}$$

on a

$$\mathbb{P}_{F_n}(S^* \leq x) = G^{n,*}(\theta + \sigma x/\sqrt{n}) = \Phi(x) + \frac{1}{n^{1/2}} \hat{p}(x)\phi(x) + \mathcal{O}_{\mathbb{P}}(\frac{1}{n}).$$

Détails

Eléments de validation asymptotique du bootstrap (3)

Le point clé est que $p(x) - \hat{p}(x) = \mathcal{O}_{\mathbb{P}}(\frac{1}{n^{1/2}})$. Un simple calcul montre alors :

- Approximation gaussienne

$$\mathbb{P}(S \leq x) - \Phi(x) = \frac{1}{n^{1/2}} p(x)\phi(x) + \mathcal{O}(\frac{1}{n}) = \mathcal{O}(\frac{1}{n^{1/2}})$$

- Approximation bootstrap

$$\mathbb{P}(S \leq x) - \mathbb{P}_{F_n}(S^* \leq x) = \mathcal{O}_{\mathbb{P}}(\frac{1}{n}).$$

Exemples

Ca marche

- pour la moyenne quand

$$\sqrt{n}\big(\bar{X}_n - \mathbb{E}(X)\big) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2)$$

- pour la médiane quand

$$\sqrt{n}\big(F_n^{-}(1/2) - F^{-}(1/2)\big) \xrightarrow{\mathcal{L}} \mathcal{N}\big(0, \frac{1}{4f^2(F^{-}(1/2))}\big)$$

Ca ne marche pas pour les extrêmes

- par exemple $X_1, \ldots, X_n$ i.i.d. $\mathcal{U}(\theta, \theta+1)$ alors $X_{(1)} \xrightarrow{\mathbb{P}} \theta$ et

$$n(X_{(1)} - \theta) \xrightarrow{\mathcal{L}} \mathcal{E}(1).$$

Intervalles de confiance par bootstrap

Intervalle de confiance du bootstrap basique

On définit les statistiques d'ordre des estimateurs bootstrap

$$\hat{\theta}^*_{(1)} \leq \hat{\theta}^*_{(2)} \leq \ldots \leq \hat{\theta}^*_{(B)}$$

IC du bootstrap basique

$$\widehat{IC}^*_{basic}(1-\alpha) = \left[2\hat{\theta} - \hat{\theta}^*_{(\lceil B(1-\alpha/2)\rceil)}, 2\hat{\theta} - \hat{\theta}^*_{(\lceil B\alpha/2\rceil)}\right]$$

Détails (1)

Détails (2)

Intervalle de confiance du percentile

S'il existe une fonction h monotone telle que la loi de $h(T)$ est symétrique autour de $h(\theta)$.

IC du percentile

$$\widehat{IC}^*_{perc}(1-\alpha) = \left[\hat{\theta}^*_{(\lceil B\alpha/2 \rceil)}, \hat{\theta}^*_{(\lceil B(1-\alpha/2) \rceil)}\right]$$

Détails (1)

Détails (2)

Intervalle de confiance du t-boostrap

Si on connaît un estimateur $\hat{\sigma} = \sigma(F_n) = \sigma(\mathcal{X})$ de la variance asymptotique $\sigma^2(F)$ de $T = \hat{\theta}$, on considère la statistique studentisée

$$S = \sqrt{n}\frac{\hat{\theta} - \theta}{\sigma(F_n)}$$

et ses versions boostrapées

$$S_b^* = \sqrt{n}\frac{\hat{\theta}_b^* - \hat{\theta}}{\sigma(\mathcal{X}_b^*)}.$$

IC du t-boostrap

$$\widehat{IC}_{tboot}^*(1-\alpha) = \left[\hat{\theta} - \frac{\sigma(F_n)}{\sqrt{n}} S^*_{(\lceil B(1-\alpha/2)\rceil)}, \hat{\theta} - \frac{\sigma(F_n)}{\sqrt{n}} S^*_{(\lceil B(\alpha/2)\rceil)}\right]$$

Détails (1)

Détails (2)

Test via l'intervalle de confiance du t-bootstrap

On considère le problème de test de $\mathcal{H}_0 : \theta = \theta_0$ v.s. $\mathcal{H}_1 : \theta \neq \theta_0$. On peut faire ce test par bootstrap en comparant la statistique de test

$$\bar{S} = |\sqrt{n}\frac{\hat{\theta} - \theta_0}{\sigma(F_n)}|$$

aux statistiques bootstrapées

$$\bar{S}_b^* = |\sqrt{n}\frac{\hat{\theta}_b^* - \hat{\theta}}{\sigma(\mathcal{X}_b^*)}|.$$

On définit alors la p-value boostrapée

$$\hat{p}_B = \frac{\#\{b : \bar{S}_b^* > \bar{S}\} + 1}{B + 1}$$

Si l'estimateur de la variance n'est pas disponible, on peut utiliser la statistique $|\hat{\theta} - \theta_0|$ et sa version bootstrappée $|\hat{\theta}_b^* - \hat{\theta}|$.

Détails