

# Bootstrap et sous-échantillonnage

Agathe Guilloux

# Plan

## Introduction

## Bootstrap

- Définitions et résultats

- Exemples d'estimateurs

- Éléments de validation asymptotique

## Intervalle de confiance par bootstrap

- Intervalle de confiance

- Test par bootstrap

# Section 1

## Introduction

## But de l'inférence statistique

On a

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

on voudrait connaître

- ▶ le biais :  $\mathbb{E}_F(T(\mathcal{X})) - \theta(F)$
- ▶ la variance :  $\mathbb{E}_F(T^2(\mathcal{X})) - \mathbb{E}_F^2(T(\mathcal{X}))$
- ▶ le MSE (EQM) :  $\mathbb{E}_F((T(\mathcal{X}) - \theta(F))^2)$
- ▶ la loi de  $T(\mathcal{X})$  :  $G(x) = \mathbb{P}_F(T(\mathcal{X}) \leq x)$ ,  $\forall x$ .
- ▶ 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  $\mathcal{X} = (X_1, \dots, X_n)$ , la fonction de répartition empirique est définie par

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

On va considérer les estimateurs par plug-in.

### Principe de plug-in

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

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

→ exemples : espérance, variance, médiane

## Section 2

# Bootstrap

## Bootstrap d'Efron [Efr82, Efr92]

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

$$\mathcal{X}_1^* = (X_{1,1}^* = X_{m_1}, \dots, X_{1,n}^* = X_{m_n})$$

...

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

...

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

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

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

# Estimateurs du bootstrap classique

Soit un paramètre inconnu  $\theta(F)$

▶ Monde réel

- ▶ avec l'échantillon initial  $\mathcal{X}$ , on définit l'estimateur  $\hat{\theta} = \theta(F_n) = T(\mathcal{X})$
- ▶ on note  $G_n$  la f.d.r. inconnue de  $\hat{\theta}$ , qui dépend de  $F$ , 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

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



## Exemples d'estimateurs bootstrap (1)

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

La f.d.r.  $G_n$  de  $\hat{\theta}$  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)

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

## Exemples d'estimateurs bootstrap (2)

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

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

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

est estimé (1ere approximation) par

$$\int (x - \int x dG_n^*)^2 dG_n^*(x)$$

puis (2ieme approximation) par

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

etc

## Éléments de validation asymptotique du bootstrap (1)

Le bootstrap fait deux approximations

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

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

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

Si  $Z_1, \dots, Z_N$  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  $B$  ?

## Éléments de validation asymptotique du bootstrap (2)

La première approximation est contrôlée par les développements d'Edgeworth (voir [?]). 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

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

Dans le monde bootstrap, on peut montrer que si

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

on a

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

## Éléments 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

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

- ▶ Approximation bootstrap

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

## Exemples

Ca marche

- ▶ pour la moyenne quand

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

- ▶ pour la médiane quand

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

Ca ne marche pas pour les extrêmes

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

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

## Section 3

### Intervalles de confiance par bootstrap

## Intervalle de confiance du bootstrap basique

On définit les statistiques d'ordre des estimateurs bootstrap

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

IC du bootstrap basique

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



## 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

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

## Test via l'intervalle de confiance du percentile

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

$$\bar{S} = |\hat{\theta} - \theta_0|$$

aux statistiques bootstrapées

$$\bar{S}_b^* = |\hat{\theta}_b^* - \hat{\theta}|.$$

On définit alors la p-value bootstrapée

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

## References



Bradley Efron, *The jackknife, the bootstrap and other resampling plans*, SIAM, 1982.



\_\_\_\_\_, *Bootstrap methods : another look at the jackknife*, Breakthroughs in Statistics, Springer, 1992, pp. 569–593.