Bootstrap

Master parcours SSD - UE Statistique Computationnelle

Pierre Mahé - bioMérieux & Université de Grenoble-Alpes

Définition

- Bootstrap : méthode d'inférence basée sur le ré-échantillonnage d'un échantillon.
- Approche non-paramétrique : pas d'hypothèse sur la loi de la variable aléatoire sous-jacente.
- Principe générique décliné pour différentes applications.

Plan

- Introduction
- Formalisation du principe de ré-échantillonnage
- Le bootstrap pour l'inférence statistique :
  - caractérisation d'un estimateur
  - intervalles de confiance
- 2 applications classiques :
  - régression
  - construction de modèles de prédiction
- Conclusion

Introduction

Principe

Cours précédent :

- méthodes de simulation pour répondre à des questions d'inférence statistique.
- données simulées selon des modèles paramétriques.

Principe

Cours précédent :

- méthodes de simulation pour répondre à des questions d'inférence statistique.
- données simulées selon des modèles paramétriques.

Le bootstrap :

- même objectif mais vise à relâcher ces hypothèses.
- se base uniquement sur le vecteur d'observations disponibles : ré-échantillonnage.

$\Rightarrow$ une approche totalement non paramétrique

Le bootstrap en deux mots

On s'intéresse à un estimateur $\hat{\theta} = \hat{\theta}(X_1, ..., X_n)$ d'un paramètre θ.

Le bootstrap en deux mots

On s'intéresse à un estimateur $\hat{\theta} = \hat{\theta}(X_1, ..., X_n)$ d'un paramètre θ.

On veut caractériser sa distribution d'échantillonnage, en se basant uniquement sur l'échantillon $(X_1, ..., X_n)$.

Le bootstrap en deux mots

On s'intéresse à un estimateur $\hat{\theta} = \hat{\theta}(X_1, ..., X_n)$ d'un paramètre θ.

On veut caractériser sa distribution d'échantillonnage, en se basant uniquement sur l'échantillon $(X_1, ..., X_n)$.

On applique la procédure suivante :

- Pour b allant de 1 à B,
- On génère un échantillon $(X_1^*, ..., X_n^*)$ en tirant avec remise dans $(X_1, ..., X_n)$.
- On calcule la valeur de notre estimateur $\hat{\theta^*}^{(b)}$ à partir de $(X_1^*, ..., X_n^*)$.

Et on travaille sur les B réalisations $\left(\hat{\theta^*}^{(b)}\right)_{b=1,...,B}$.

Le bootstrap en deux mots

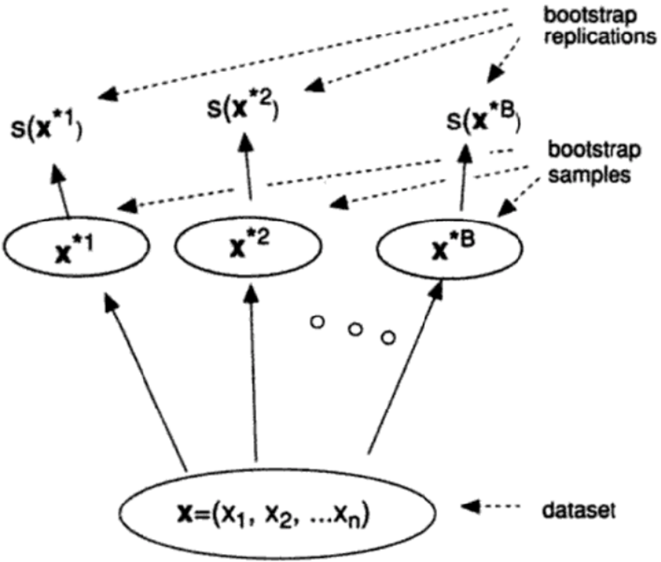
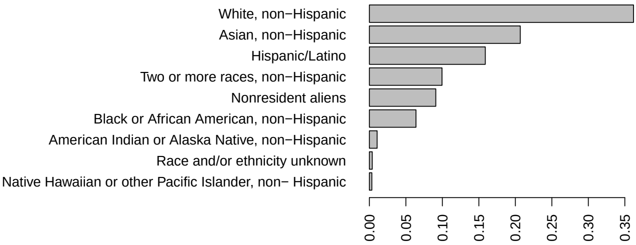


Illustration [1]

Données : origine ethnique des étudiants de Standford



- On connaît toute la population
- Elle contient 6.4% d'étudiants afro-américains.

⇒ Question : estimer ce taux à partir d'un échantillon.

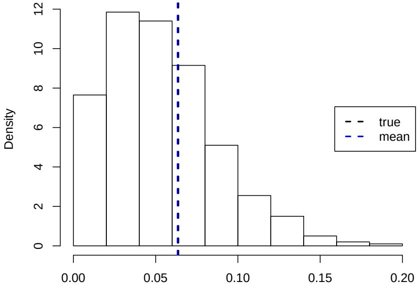
1. Tirée de https://web.stanford.edu/class/stats101/sampling/sampling-lecture02.pdf

Illustration

On considère :

- des échantillons de taille 50.
- l'estimateur de la fréquence empirique.

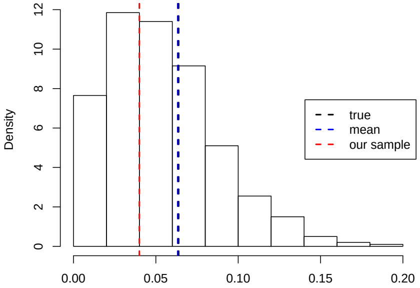
⇒ distribution d'échantillonnage sur 1000 réalisations :



⇒ estimateur précis en moyenne, mais forte variance.

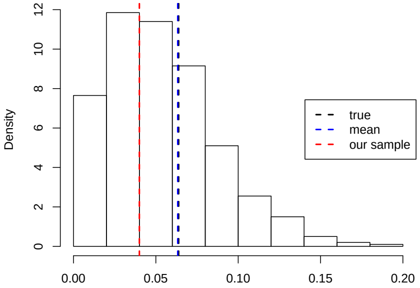
Illustration

Sauf qu'en pratique, on n'aurait accès qu'à une réalisation :



Illustration

Sauf qu'en pratique, on n'aurait accès qu'à une réalisation :

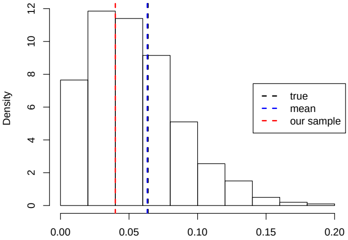


⇒ bootstrap : se baser uniquement sur notre échantillon :

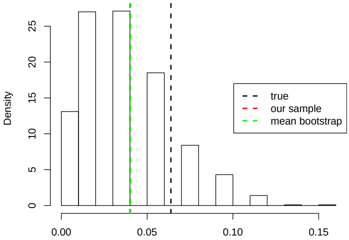
- nombreux tirages avec remise
- distribution de la fréquence empirique "bootstrapée"

Illustration

Tirages dans la population :



Tirages dans l'échantillon :



- l'échantillon $\sim$ une population finie
- on simule des échantillons de cette population
- on veut en tirer des conclusions sur la vraie population

Illustration

Remarques :

- intérêt limité sur cet exemple, car on connaît très bien les propriétés de l'estimateur de la fréquence empirique.
- intérêt général de la démarche :
  - ne pas "se forcer" à faire d'hypothèses quand les données ne s'y prètent pas
  - permet de considérer des statistiques plus complexes
    - dont on ne connaît pas forcément la distribution d'échantillonnage

Bootstrap, vous avez dit bootstrap ?

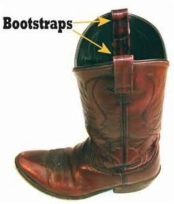
Terme introduit en statistique par Bradley Efron en 1979...

... vient de l'expression "to pull oneself up by one's bootstrap"...

- $\sim$ s'en sortir par soi même, grâce à ses efforts

...souvent attribuée aux aventures du Baron de Munchausen.

- le Baron, tombé dans un marécage, s'en extrait en se tirant lui même par ses "bootstraps"



Formalisation du principe de ré-échantillonnage

Récapitulatif

On s'intéresse à une variable aléatoire X.

Récapitulatif

On s'intéresse à une variable aléatoire X.

Au niveau de la population $\mathcal{X}$:

- X est régie par une distribution P et une fonction de répartition F :

$$P(X \leq x) = F(x), \;\; \forall x \in \mathcal{X}.$$

- $\theta = t(P)$ est un paramètre d'intérêt

Récapitulatif

On s'intéresse à une variable aléatoire X.

Au niveau de la population $\mathcal{X}$:

- X est régie par une distribution P et une fonction de répartition F :

$$P(X \leq x) = F(x), \ \ \forall x \in \mathcal{X}.$$

- $\theta = t(P)$ est un paramètre d'intérêt

On dispose d'un échantillon $\mathbf{X} = (X_1, ..., X_n)$:

- les X_i sont iid selon P
- $\hat{\theta} = s(\mathbf{X})$ est un estimateur de θ

Récapitulatif

On s'intéresse à une variable aléatoire X.

Au niveau de la population $\mathcal{X}$:

- X est régie par une distribution P et une fonction de répartition F :

$$P(X \leq x) = F(x), \;\; \forall x \in \mathcal{X}.$$

- $\theta = t(P)$ est un paramètre d'intérêt

On dispose d'un échantillon $\mathbf{X} = (X_1, ..., X_n)$:

- les X_i sont iid selon P
- $\hat{\theta} = s(\mathbf{X})$ est un estimateur de θ

$\Rightarrow$ inférence : tirer des conclusions sur θ à partir de $\hat{\theta}$.
$\Rightarrow$ nécessite la distribution d'échantillonnage de $\hat{\theta}$.

Récapitulatif

Question clé : estimer la distribution d'échantillonnage de $\hat{\theta}$.

Récapitulatif

Question clé : estimer la distribution d'échantillonnage de $\hat{\theta}$.

Stratégie #1 : l'estimer empiriquement à partir de plusieurs échantillons :

- on collecte plusieurs échantillons $\mathbf{X_j} = (X_{j1}, ..., X_{jn})$
- on calcule $\hat{\theta}_j = s(\mathbf{X_j})$
- on l'estime par la distribution des $\{\hat{\theta}_j\}$

Récapitulatif

Question clé : estimer la distribution d'échantillonnage de $\hat{\theta}$.

Stratégie #1 : l'estimer empiriquement à partir de plusieurs échantillons :

- on collecte plusieurs échantillons $\mathbf{X_j} = (X_{j1}, ..., X_{jn})$
- on calcule $\hat{\theta}_j = s(\mathbf{X_j})$
- on l'estime par la distribution des $\{\hat{\theta}_j\}$

Mais en pratique on ne dispose que d'un échantillon $\mathbf{X}$...

Récapitulatif

Question clé : estimer la distribution d'échantillonnage de $\hat{\theta}$.

Stratégie #1 : l'estimer empiriquement à partir de plusieurs échantillons :

- on collecte plusieurs échantillons $\mathbf{X_j} = (X_{j1}, ..., X_{jn})$
- on calcule $\hat{\theta}_j = s(\mathbf{X_j})$
- on l'estime par la distribution des $\{\hat{\theta}_j\}$

Mais en pratique on ne dispose que d'un échantillon $\mathbf{X}$...

$\Rightarrow$ approche paramétrique : faire des hypothèses sur la nature de la distribution P des données.

$\Rightarrow$ approche non-paramétrique : ré-échantillonnage dans $\mathbf{X}$ pour estimer la distribution d'échantillonnage de $\hat{\theta}$.

Distribution empirique

Soit un échantillon $\mathbf{X} = (X_1, ..., X_n)$.

On définit la distribution empirique $\hat{P}_n$ comme :

$$\hat{P}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(X_i = x).$$

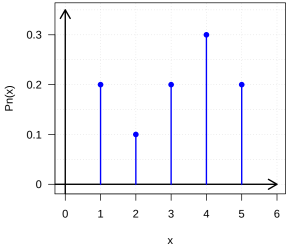
On définit de même la fonction de répartition empirique $\hat{F}_n$:

$$\hat{F}_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(X_i \leq t).$$

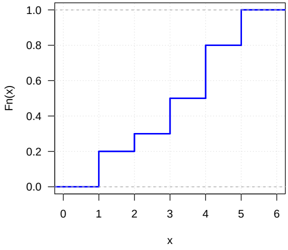
Distribution empirique - illustration

Soit le vecteur $x = \{1, 1, 2, 3, 3, 4, 4, 4, 5, 5\}$.

$\Rightarrow$ distribution empirique $\hat{P}_n$:

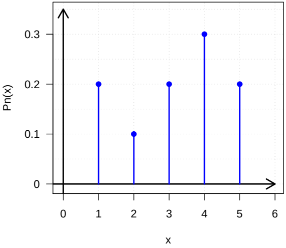


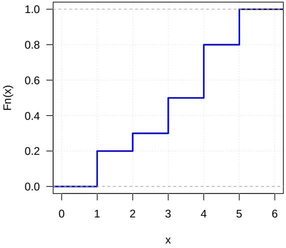
$\Rightarrow$ répartition empirique $\hat{F}_n$:



Distribution empirique & ré-échantillonnage

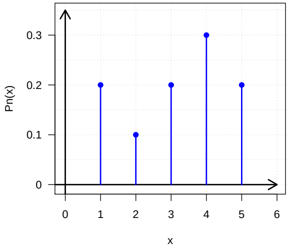
Comment simuler un échantillon selon une distribution empirique ?

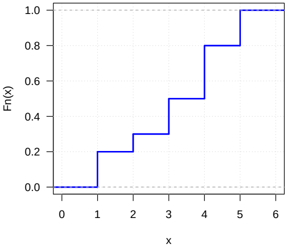




Distribution empirique & ré-échantillonnage

Comment simuler un échantillon selon une distribution empirique ?





$\Rightarrow$ il suffit de tirer avec remise dans l'échantillon original.

$\Rightarrow$ pour s'en convaincre, on peut utiliser la méthode d'inversion.

Ré-échantillonnage & bootstrap

On s'intéresse à un estimateur $\hat{\theta} = \hat{\theta}(X_1, ..., X_n)$ d'un paramètre θ.

Ré-échantillonnage & bootstrap

On s'intéresse à un estimateur $\hat{\theta} = \hat{\theta}(X_1, ..., X_n)$ d'un paramètre θ.

On veut caractériser sa distribution d'échantillonnage, en se basant uniquement sur l'échantillon $(X_1, ..., X_n)$.

Ré-échantillonnage & bootstrap

On s'intéresse à un estimateur $\hat{\theta} = \hat{\theta}(X_1, ..., X_n)$ d'un paramètre θ.

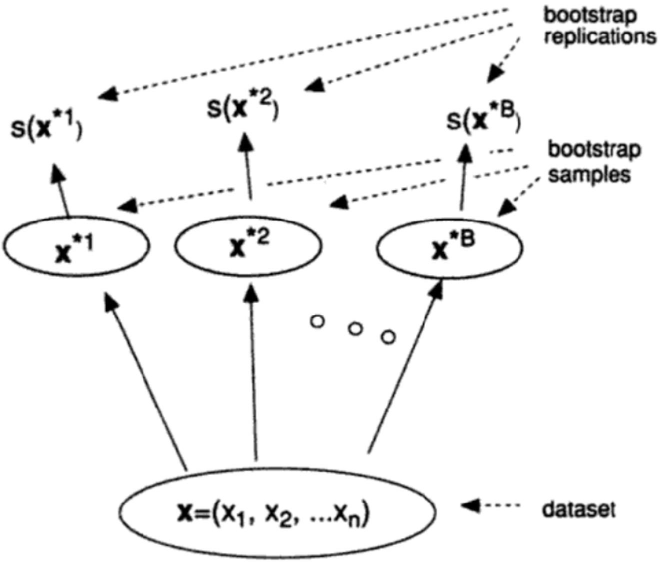
On veut caractériser sa distribution d'échantillonnage, en se basant uniquement sur l'échantillon $(X_1, ..., X_n)$.

On applique la procédure suivante :

- Pour b allant de 1 à B,
- On génère un échantillon $(X_1^*, ..., X_n^*)$ en tirant avec remise dans $(X_1, ..., X_n)$.
- On calcule la valeur de notre estimateur $\hat{\theta^*}^{(b)}$ à partir de $(X_1^*, ..., X_n^*)$.

Et on travaille sur les B réalisations $\left(\hat{\theta^*}^{(b)}\right)_{b=1,...,B}$.

Ré-échantillonnage & bootstrap

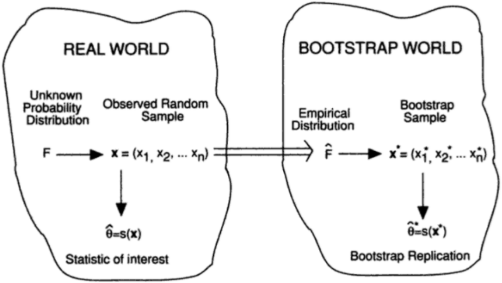


Monde réel et monde bootstrap

Formellement, le bootstrap considère des réalisations tirées de la distribution empirique définie par l'échantillon original.

On passe du monde réel au "monde bootstrap" :

- illustration tirée de Efron and Tibshirani (1993).



Monde réel et monde bootstrap

Monde réel vs monde bootstrap :

Monde réel et monde bootstrap

Monde réel vs monde bootstrap :

- distribution des données : P vs $\hat{P}_n$
  - vraie distribution P inconnue
  - distribution $\hat{P}_n$ parfaitement connue

Monde réel et monde bootstrap

Monde réel vs monde bootstrap :

- distribution des données : P vs $\hat{P}_n$
  - vraie distribution P inconnue
  - distribution $\hat{P}_n$ parfaitement connue
- échantillon : $(X_1, ..., X_n)$ vs $(X_1^*, ..., X_n^*)$
  - un unique échantillon $(X_1, ..., X_n) \sim P$
  - autant d'échantillons $(X_1^*, ..., X_n^*) \sim \hat{P}_n$ qu'on veut

Monde réel et monde bootstrap

Monde réel vs monde bootstrap :

- distribution des données : P vs $\hat{P}_n$
  - vraie distribution P inconnue
  - distribution $\hat{P}_n$ parfaitement connue
- échantillon : $(X_1, ..., X_n)$ vs $(X_1^*, ..., X_n^*)$
  - un unique échantillon $(X_1, ..., X_n) \sim P$
  - autant d'échantillons $(X_1^*, ..., X_n^*) \sim \hat{P}_n$ qu'on veut
- paramètre : θ vs $\hat{\theta}$ vs $\hat{\theta}^*$
  - un vrai paramètre θ inconnu
  - une estimation $\hat{\theta} = s(X_1, ..., X_n)$ connue
    - NB : une estimation de θ
  - autant d'estimations $\hat{\theta}^* = s(X_1^*, ..., X_n^*)$ qu'on veut
    - NB : des estimations de $\hat{\theta}$

Monde réel et monde bootstrap

Dans le monde bootstrap :

- la distribution des données est $\hat{P}_n$
  - elle est **parfaitement connue**
  - on peut en **tirer/simuler des échantillons**
- le "vrai" paramètre de la population est $\hat{\theta}$
  - il est **parfaitement connu**
  - on peut le **comparer aux estimations** $\hat{\theta}^*$

Monde réel et monde bootstrap

Dans le monde bootstrap :

- la distribution des données est $\hat{P}_n$
  - elle est **parfaitement connue**
  - on peut en **tirer/simuler des échantillons**
- le "vrai" paramètre de la population est $\hat{\theta}$
  - il est **parfaitement connu**
  - on peut le **comparer aux estimations** $\hat{\theta}^*$

$\Rightarrow$ principe du bootstrap :

1. se placer dans le monde bootstrap
2. calculer la distribution d'échantillonnage des $\hat{\theta}^*$
3. se comparer à $\hat{\theta}$ (le "vrai" paramètre de $\hat{P}_n$)
4. en déduire des caractéristiques de $\hat{\theta}$ (par rapport à P).
   - e.g., biais, erreur-type et intervalle de confiance

Bootstrap et inférence statistique : caractérisation d'un estimateur

Bootstrap et estimation

On peut par exemple appliquer le bootstrap pour :

- Estimer le biais d'un estimateur.
- Estimer son erreur quadratique moyenne.
- Estimer son erreur type.
- Donner un intervalle de confiance sur une estimation.

Estimer le biais d'un estimateur

Rappel : biais d'un estimateur : $\text{Biais}(\hat{\theta}) = E\left[\hat{\theta}\right] - \theta$.

Estimer le biais d'un estimateur

Rappel : biais d'un estimateur : $\text{Biais}(\hat{\theta}) = E\left[\hat{\theta}\right] - \theta$.

Procédure bootstrap :

1. On dispose d'un échantillon $(X_1, ..., X_n)$.
2. On calcule $\hat{\theta}$ sur l'échantillon.
3. On applique la procédure bootstrap pour obtenir des réalisations $\hat{\theta^*}^{(b)}$:
   - Pour b allant de 1 à B,
   - on génère un échantillon $(X_1^*, ..., X_n^*)$,
   - on calcule $\hat{\theta^*}^{(b)}$ à partir de $(X_1^*, ..., X_n^*)$.

Estimer le biais d'un estimateur

Rappel : biais d'un estimateur : $\text{Biais}(\hat{\theta}) = E\left[\hat{\theta}\right] - \theta$.

Procédure bootstrap :

1. On dispose d'un échantillon $(X_1, ..., X_n)$.
2. On calcule $\hat{\theta}$ sur l'échantillon.
3. On applique la procédure bootstrap pour obtenir des réalisations $\hat{\theta^*}^{(b)}$:
   - Pour b allant de 1 à B,
   - on génère un échantillon $(X_1^*, ..., X_n^*)$,
   - on calcule $\hat{\theta^*}^{(b)}$ à partir de $(X_1^*, ..., X_n^*)$.

On estime le biais de $\hat{\theta}$ par :

$$\widehat{\text{Biais}(\hat{\theta})} = \frac{1}{B}\sum_{b=1}^{B} \hat{\theta^*}^{(b)} - \hat{\theta}.$$

Estimer l'erreur quadratique moyenne (MSE) d'un estimateur

Rappel : MSE d'un estimateur : $\text{MSE}(\hat{\theta}) = E\left[(\hat{\theta} - \theta)^2\right]$.

Estimer l'erreur quadratique moyenne (MSE) d'un estimateur

Rappel : MSE d'un estimateur : $\text{MSE}(\hat{\theta}) = E\left[(\hat{\theta} - \theta)^2\right]$.

Procédure bootstrap :

1. On dispose d'un échantillon $(X_1, ..., X_n)$.
2. On calcule $\hat{\theta}$ sur l'échantillon.
3. On applique la procédure bootstrap pour obtenir des réalisations $\hat{\theta^*}^{(b)}$:
   - Pour b allant de 1 à B,
   - on génère un échantillon $(X_1^*, ..., X_n^*)$,
   - on calcule $\hat{\theta^*}^{(b)}$ à partir de $(X_1^*, ..., X_n^*)$.

Estimer l'erreur quadratique moyenne (MSE) d'un estimateur

Rappel : MSE d'un estimateur : $\text{MSE}(\hat{\theta}) = E\left[(\hat{\theta} - \theta)^2\right]$.

Procédure bootstrap :

1. On dispose d'un échantillon $(X_1, ..., X_n)$.
2. On calcule $\hat{\theta}$ sur l'échantillon.
3. On applique la procédure bootstrap pour obtenir des réalisations $\hat{\theta^*}^{(b)}$:
   - Pour b allant de 1 à B,
   - on génère un échantillon $(X_1^*, ..., X_n^*)$,
   - on calcule $\hat{\theta^*}^{(b)}$ à partir de $(X_1^*, ..., X_n^*)$.

On estime l'erreur quadratique moyenne de $\hat{\theta}$ par :

$$\widehat{\text{MSE}(\hat{\theta})} = \frac{1}{B} \sum_{b=1}^{B} \left(\hat{\theta^*}^{(b)} - \hat{\theta}\right)^2.$$

Estimer l'erreur type d'un estimateur

Rappel : erreur type d'un estimateur : l'écart type de sa distribution d'échantillonnage.

Estimer l'erreur type d'un estimateur

Rappel : erreur type d'un estimateur : l'écart type de sa distribution d'échantillonnage.

Procédure bootstrap :

1. On dispose d'un échantillon $(X_1, ..., X_n)$.
2. On applique la procédure bootstrap pour obtenir des réalisations $\hat{\theta^*}^{(b)}$:
   - Pour b allant de 1 à B,
   - on génère un échantillon $(X_1^*, ..., X_n^*)$,
   - on calcule $\hat{\theta^*}^{(b)}$ à partir de $(X_1^*, ..., X_n^*)$.

Estimer l'erreur type d'un estimateur

Rappel : erreur type d'un estimateur : l'écart type de sa distribution d'échantillonnage.

Procédure bootstrap :

1. On dispose d'un échantillon $(X_1, ..., X_n)$.
2. On applique la procédure bootstrap pour obtenir des réalisations $\hat{\theta^*}^{(b)}$:
   - Pour b allant de 1 à B,
   - on génère un échantillon $(X_1^*, ..., X_n^*)$,
   - on calcule $\hat{\theta^*}^{(b)}$ à partir de $(X_1^*, ..., X_n^*)$.

On estime l'erreur-type de $\hat{\theta}$ par l'écart-type des $\left(\hat{\theta^*}^{(b)}\right)$:

$$\widehat{\mathrm{se}(\hat{\theta})} = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\theta^*}^{(b)} - \bar{\hat{\theta^*}}\right)^2} \text{ où } \bar{\hat{\theta^*}} = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta^*}^{(b)}.$$

Remarque importante

In fine, on estime la "vraie" variance de $\hat{\theta}$ (i.e., selon P) par la variance empirique des $\left(\hat{\theta}^{*(b)}\right)$ (i.e., selon $\hat{P}_n$).

Remarque importante

In fine, on estime la "vraie" variance de $\hat{\theta}$ (i.e., selon P) par la variance empirique des $\left(\hat{\theta^*}^{(b)}\right)$ (i.e., selon $\hat{P}_n$).

Il faut bien comprendre qu'il y a 2 niveaux d'approximation :

1. approximer la vraie distribution P par l'empirique $\hat{P}_n$.
2. approximer la variance de $\hat{\theta}$ selon $\hat{P}_n$ par la variance empirique des $\left(\hat{\theta^*}^{(b)}\right)$.

Remarque importante

In fine, on estime la "vraie" variance de $\hat{\theta}$ (i.e., selon P) par la variance empirique des $\left(\hat{\theta^*}^{(b)}\right)$ (i.e., selon $\hat{P}_n$).

Il faut bien comprendre qu'il y a 2 niveaux d'approximation :

1. approximer la vraie distribution P par l'empirique $\hat{P}_n$.
2. approximer la variance de $\hat{\theta}$ selon $\hat{P}_n$ par la variance empirique des $\left(\hat{\theta^*}^{(b)}\right)$.

$\Rightarrow$ point #2 : ok, il suffit de prendre B grand.
$\Rightarrow$ point #1 : plus délicat...mais valide quand n est grand.

Remarque importante

In fine, on estime la "vraie" variance de $\hat{\theta}$ (i.e., selon P) par la variance empirique des $(\hat{\theta^*}^{(b)})$ (i.e., selon $\hat{P}_n$).

Il faut bien comprendre qu'il y a 2 niveaux d'approximation :

1. approximer la vraie distribution P par l'empirique $\hat{P}_n$.
2. approximer la variance de $\hat{\theta}$ selon $\hat{P}_n$ par la variance empirique des $(\hat{\theta^*}^{(b)})$.

$\Rightarrow$ point #2 : ok, il suffit de prendre B grand.
$\Rightarrow$ point #1 : plus délicat...mais valide quand n est grand.

⚠ bootstrap $\neq$ méthode pour petits échantillons !

- intérêt #1 = relâcher hypothèses paramétriques
- intérêt #2 = générique, valable pour toute statistique

Illustration

Calcul du biais et de l'erreur type de la moyenne :

```
> x = c(1,1,2,3,3,4,4,4,5,5)
> n = length(x)
> B = 2000
> theta.hat = mean(x)
> theta.b = numeric(B)
> for(b in 1:B){
ind = sample(1:n, size = n, replace = TRUE)
theta.b[b] = mean(x[ind])
}
> bias = mean(theta.b) - theta.hat
> se = sd(theta.b)
```

Bootstrap et inférence statistique : intervalles de confiance

Bootstrap et intervalles de confiance

Il existe de nombreuses manières de définir des intervalles de confiance par bootstrap.

Bootstrap et intervalles de confiance

Il existe de nombreuses manières de définir des intervalles de confiance par bootstrap.

Nous allons considérer deux définitions se basant uniquement sur les quantiles de la distribution des $\hat{\theta}^{*(b)}$:

- l'intervalle de confiance des percentiles.
- l'intervalle de confiance basique (ou "du pivot").

Ces définitions font le moins d'hypothèses possible.

Intervalle de confiance - méthode des percentiles

Cette définition est probablement la plus simple et intuitive.

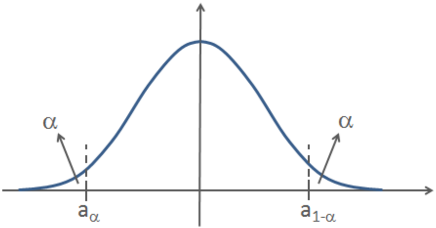
Elle consiste à calculer empiriquement l'intervalle couvrant $100(1-\alpha)\%$ des estimations bootstrap obtenues.

Formellement, en notant q^*_α le quantile d'ordre α de la distribution des estimations bootstrap $\hat{\theta}^{*(b)}$, il est défini comme :

$$\left[q^*_{\alpha/2} \;\; ; \;\; q^*_{1-\alpha/2}\right].$$

Intervalle de confiance - méthode basique (1/3)

On va s'intéresser à la distribution de la statistique $(\hat{\theta} - \theta)$:



Si a_α le quantile d'ordre α de la statistique $(\hat{\theta} - \theta)$ alors :

$$P(\hat{\theta} - \theta \leq a_\alpha) = \alpha \text{ et } P(\hat{\theta} - \theta \geq a_{1-\alpha}) = \alpha.$$

Intervalle de confiance - méthode basique (2/3)

On a donc (par définition) :

$$P(\hat{\theta} - \theta \leq a_\alpha) = \alpha \text{ et } P(\hat{\theta} - \theta \geq a_{1-\alpha}) = \alpha,$$

où a_α le quantile d'ordre α de la statistique $(\hat{\theta} - \theta)$.

Intervalle de confiance - méthode basique (2/3)

On a donc (par définition) :

$$P(\hat{\theta} - \theta \leq a_\alpha) = \alpha \text{ et } P(\hat{\theta} - \theta \geq a_{1-\alpha}) = \alpha,$$

où a_α le quantile d'ordre α de la statistique $(\hat{\theta} - \theta)$.

$\Rightarrow$ on peut écrire :

$$P(\theta \geq \hat{\theta} - a_\alpha) = \alpha \text{ et } P(\theta \leq \hat{\theta} - a_{1-\alpha}) = \alpha,$$

Intervalle de confiance - méthode basique (2/3)

On a donc (par définition) :

$$P(\hat{\theta} - \theta \leq a_\alpha) = \alpha \text{ et } P(\hat{\theta} - \theta \geq a_{1-\alpha}) = \alpha,$$

où a_α le quantile d'ordre α de la statistique $(\hat{\theta} - \theta)$.

$\Rightarrow$ on peut écrire :

$$P(\theta \geq \hat{\theta} - a_\alpha) = \alpha \text{ et } P(\theta \leq \hat{\theta} - a_{1-\alpha}) = \alpha,$$

et en déduire l'intervalle de confiance à $100(1-\alpha)\%$:

$$\left[\hat{\theta} - a_{1-\alpha/2} \ ; \ \hat{\theta} - a_{\alpha/2}\right].$$

Intervalle de confiance - méthode basique (3/3)

On considère donc l'intervalle de confiance défini comme

$$\left[\hat{\theta} - a_{1-\alpha/2} \;\; ; \;\; \hat{\theta} - a_{\alpha/2}\right],$$

où a_α le quantile d'ordre α de la statistique $(\hat{\theta} - \theta)$.

Intervalle de confiance - méthode basique (3/3)

On considère donc l'intervalle de confiance défini comme

$$\left[\hat{\theta} - a_{1-\alpha/2} \;\; ; \;\; \hat{\theta} - a_{\alpha/2}\right],$$

où a_α le quantile d'ordre α de la statistique $(\hat{\theta} - \theta)$.

Problème : on ne connaît pas la distribution de $(\hat{\theta} - \theta)$.

Intervalle de confiance - méthode basique (3/3)

On considère donc l'intervalle de confiance défini comme

$$\left[\hat{\theta} - a_{1-\alpha/2} \;\; ; \;\; \hat{\theta} - a_{\alpha/2}\right],$$

où a_α le quantile d'ordre α de la statistique $(\hat{\theta} - \theta)$.

Problème : on ne connaît pas la distribution de $(\hat{\theta} - \theta)$.

$\Rightarrow$ On passe dans le monde bootstrap : on approxime la distribution de $(\hat{\theta} - \theta)$ par celle de $(\hat{\theta^*}^{(b)} - \hat{\theta})$.

Intervalle de confiance - méthode basique (3/3)

On considère donc l'intervalle de confiance défini comme

$$\left[\hat{\theta} - a_{1-\alpha/2} \;\; ; \;\; \hat{\theta} - a_{\alpha/2}\right],$$

où a_α le quantile d'ordre α de la statistique $(\hat{\theta} - \theta)$.

Problème : on ne connaît pas la distribution de $(\hat{\theta} - \theta)$.

$\Rightarrow$ On passe dans le monde bootstrap : on approxime la distribution de $(\hat{\theta} - \theta)$ par celle de $(\hat{\theta^*}^{(b)} - \hat{\theta})$.

- on peut donc écrire :
  $a_\alpha = q^*_\alpha - \hat{\theta}$, où q^*_α est le quantile d'ordre α des $\hat{\theta^*}^{(b)}$.

Intervalle de confiance - méthode basique (3/3)

On considère donc l'intervalle de confiance défini comme

$$\left[\hat{\theta} - a_{1-\alpha/2} \;\; ; \;\; \hat{\theta} - a_{\alpha/2}\right],$$

où a_α le quantile d'ordre α de la statistique $(\hat{\theta} - \theta)$.

Problème : on ne connaît pas la distribution de $(\hat{\theta} - \theta)$.

$\Rightarrow$ On passe dans le monde bootstrap : on approxime la distribution de $(\hat{\theta} - \theta)$ par celle de $(\hat{\theta^*}^{(b)} - \hat{\theta})$.

- on peut donc écrire :
  $a_\alpha = q^*_\alpha - \hat{\theta}$, où q^*_α est le quantile d'ordre α des $\hat{\theta^*}^{(b)}$.
- on en déduit la définition suivante :

$$\left[2\hat{\theta} - q^*_{1-\alpha/2} \;\; ; \;\; 2\hat{\theta} - q^*_{\alpha/2}\right].$$

Bootstrap & intervalles de confiance - remarques

Deux définitions considérées :

1. IC des percentiles : $\left[q^*_{\alpha/2}\ ;\ q^*_{1-\alpha/2}\right]$.
2. IC basique : $\left[2\hat{\theta} - q^*_{1-\alpha/2}\ ;\ 2\hat{\theta} - q^*_{\alpha/2}\right]$.

$\Rightarrow$ se basent uniquement sur q^*_α : le quantile d'ordre α des estimations bootstrap $\hat{\theta}^{*(b)}$.

$\Rightarrow$ méthode "basique" / "pivot" : analogie monde bootstrap

- $\theta \to \hat{\theta}\,;\ \hat{\theta} \to \hat{\theta}^*$

Bootstrap & intervalles de confiance - remarques

Deux définitions considérées :

1. IC des percentiles : $\left[q^*_{\alpha/2} \ ; \ q^*_{1-\alpha/2}\right]$.
2. IC basique : $\left[2\hat{\theta} - q^*_{1-\alpha/2} \ ; \ 2\hat{\theta} - q^*_{\alpha/2}\right]$.

⇒ se basent uniquement sur q^*_α : le quantile d'ordre α des estimations bootstrap $\hat{\theta^*}^{(b)}$.

⇒ méthode "basique" / "pivot" : analogie monde bootstrap

- $\theta \to \hat{\theta}\,;\, \hat{\theta} \to \hat{\theta}^*$

De nombreuses autres définitions existent.

- normal, "studentisé", accéléré, corrigé du biais...

Applications : bootstrap et régression

Régression linéaire

Objectif : prédire / modéliser une variable $Y \in \mathbb{R}$ à partir de p variables explicatives $X^j \in \mathbb{R}$.

Modèle :

$$\boxed{Y = \beta_0 + \sum_{j=1}^{p} \beta_j X^j + \epsilon}$$ avec ϵ un terme d'erreur.

Régression linéaire

Objectif : prédire / modéliser une variable $Y \in \mathbb{R}$ à partir de p variables explicatives $X^j \in \mathbb{R}$.

Modèle :

$$\boxed{Y = \beta_0 + \sum_{j=1}^{p} \beta_j X^j + \epsilon}$$ avec ϵ un terme d'erreur.

$\Rightarrow$ on estime les coefficients β_j par moindre carrés

- à partir d'un échantillon $(X_i, Y_i) \in \mathbb{R}^p \times \mathbb{R}$, $i = 1, ..., n$

Régression linéaire

Objectif : prédire / modéliser une variable $Y \in \mathbb{R}$ à partir de p variables explicatives $X^j \in \mathbb{R}$.

Modèle :

$$\boxed{Y = \beta_0 + \sum_{j=1}^{p} \beta_j X^j + \epsilon}$$ avec ϵ un terme d'erreur.

$\Rightarrow$ on estime les coefficients β_j par moindre carrés

- à partir d'un échantillon $(X_i, Y_i) \in \mathbb{R}^p \times \mathbb{R}$, $i = 1, ..., n$

$\Rightarrow$ sous l'hypothèse que les résidus ϵ_i sont iid selon $\mathcal{N}(0, \sigma^2)$ on connait la distribution d'échantillonnage des $\hat{\beta}_j$.

- on peut donc en tirer des intervalles de confiance

Approche bootstrap

Bootstrap et régression = alternative non-paramétrique.

- relâcher les hypotèses du modèle.

Objectif : estimer la distribution d'échantillonnage des $\hat{\beta}_j$.

- in fine : calcul d'intervalles de confiance comme avant.

Deux stratégies principales :

1. bootstrap par paires
2. bootstrap des résidus

Bootstrap par paires

Principe : tirer avec remise dans $\{(X_i, Y_i)\}_{i=1,...,n}$.

Procédure bootstrap :

- On dispose d'un échantillon $(Z_1, ..., Z_n)$, $Z_i = (X_i, Y_i)$
- On estime $\hat{\beta}_j$ sur l'échantillon original
- On applique le bootstrap par paires :
  - pour b de 1 à B
  - on génère un échantillon $(Z_1^*, ..., Z_n^*)$
  - on estime les coefficients $\hat{\beta}_j^*$ à partir des $(Z_1^*, ..., Z_n^*)$

Bootstrap par paires

Principe : tirer avec remise dans $\{(X_i, Y_i)\}_{i=1,...,n}$.

Procédure bootstrap :

- On dispose d'un échantillon $(Z_1, ..., Z_n)$, $Z_i = (X_i, Y_i)$
- On estime $\hat{\beta}_j$ sur l'échantillon original
- On applique le bootstrap par paires :
  - pour b de 1 à B
  - on génère un échantillon $(Z_1^*, ..., Z_n^*)$
  - on estime les coefficients $\hat{\beta}_j^*$ à partir des $(Z_1^*, ..., Z_n^*)$

$\Rightarrow$ l'approche "standard".

$\Rightarrow$ hypothèse : (X_i, Y_i) iid selon une loi (jointe) P.

Bootstrap des résidus

Principe : travailler à partir des résidus du modèle initial.

Procédure bootstrap :

- On dispose d'un échantillon $(Z_1, ..., Z_n)$, $Z_i = (X_i, Y_i)$
- On estime $\hat{\beta}_j$ sur l'échantillon original
- On considère les résidus $(\epsilon_1, ..., \epsilon_n)$
- On applique le bootstrap par résidus :
  - pour b de 1 à B
  - on génère un échantillon $(\epsilon_1^*, ..., \epsilon_n^*)$
  - on calcule $Y_i^* = \hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j X_{ij} + \epsilon_i^*$
  - on estime les coefficients $\hat{\beta}_j^*$ à partir des (X_i, Y_i^*)

Bootstrap des résidus

Principe : travailler à partir des résidus du modèle initial.

Procédure bootstrap :

- On dispose d'un échantillon $(Z_1, ..., Z_n)$, $Z_i = (X_i, Y_i)$
- On estime $\hat{\beta}_j$ sur l'échantillon original
- On considère les résidus $(\epsilon_1, ..., \epsilon_n)$
- On applique le bootstrap par résidus :
  - pour b de 1 à B
  - on génère un échantillon $(\epsilon_1^*, ..., \epsilon_n^*)$
  - on calcule $Y_i^* = \hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j X_{ij} + \epsilon_i^*$
  - on estime les coefficients $\hat{\beta}_j^*$ à partir des (X_i, Y_i^*)

$\Rightarrow$ on tire avec remise les résidus.

- tous les X_i sont utilisés à chaque fois
- hypothèse : $Y_i | X_i$ iid

$\Rightarrow$ semble être la stratégie la plus classique

Applications : bootstrap et modèles de prédiction

Stratégie bootstrap & "bagging"

Principe bootstrap pour la construction de prédicteurs :

Stratégie bootstrap & "bagging"

Principe bootstrap pour la construction de prédicteurs :

1. Apprentissage : construire B prédicteurs à partir d'échantillons obtenus en tirant avec remise dans l'échantillon original.

Stratégie bootstrap & "bagging"

Principe bootstrap pour la construction de prédicteurs :

1. Apprentissage : construire B prédicteurs à partir d'échantillons obtenus en tirant avec remise dans l'échantillon original.
2. Prédiction : aggréger les prédictions des B modèles
   - régression : prédire la moyenne des valeurs obtenues.
   - classification : prédire la classe prédite le plus souvent

Stratégie bootstrap & "bagging"

Principe bootstrap pour la construction de prédicteurs :

1. Apprentissage : construire B prédicteurs à partir d'échantillons obtenus en tirant avec remise dans l'échantillon original.
2. Prédiction : aggréger les prédictions des B modèles
   - régression : prédire la moyenne des valeurs obtenues.
   - classification : prédire la classe prédite le plus souvent

On parle de stratégie bagging, pour bootstrap-aggregating.

Stratégie générique, souvent basée sur des arbres de décision

En pratique elle permet de limiter le sur-apprentissage.

Stratégie bootstrap & "bagging"

Principe bootstrap pour la construction de prédicteurs :

1. Apprentissage : construire B prédicteurs à partir d'échantillons obtenus en tirant avec remise dans l'échantillon original.
2. Prédiction : aggréger les prédictions des B modèles
   - régression : prédire la moyenne des valeurs obtenues.
   - classification : prédire la classe prédite le plus souvent

On parle de stratégie bagging, pour bootstrap-aggregating.

Stratégie générique, souvent basée sur des arbres de décision

En pratique elle permet de limiter le sur-apprentissage.

$\Rightarrow$ à suivre dans cours "Apprentissage Statistique II".

Conclusion

Remarques / Conclusion

- Bootstrap : un principe très simple à mettre en oeuvre.

Remarques / Conclusion

- Bootstrap : un principe très simple à mettre en oeuvre.
- Il permet de répondre à des questions d'inférence statistique sans aucune information sur la loi de la variable aléatoire sous-jacente.

Remarques / Conclusion

- Bootstrap : un principe très simple à mettre en oeuvre.
- Il permet de répondre à des questions d'inférence statistique sans aucune information sur la loi de la variable aléatoire sous-jacente.
- Dans ce cas là, travailler par ré-échantillonnage de l'échantillon disponible est parfois la meilleure stratégie, surtout si la loi sous-jacente n'est pas une loi usuelle.

Remarques / Conclusions

- Intérêts principaux :
  1. relâcher les hypothèses sur la loi de la variable aléatoire étudiée qu'on doit faire avec les approches paramétriques
  2. principe générique applicable à n'importe quelle statistique (dont on ne connait pas la distribution)

Remarques / Conclusions

- Intérêts principaux :
  1. relâcher les hypothèses sur la loi de la variable aléatoire étudiée qu'on doit faire avec les approches paramétriques
  2. principe générique applicable à n'importe quelle statistique (dont on ne connait pas la distribution)

- Bien garder en tête qu'il y a 2 niveaux d'approximation
  1. remplacer la "vraie" distribution par l'empirique
  2. remplacer la "vraie" distribution d'échantillonnage (selon la loi empirique) par celle obtenue par tirages

Remarques / Conclusions

- Intérêts principaux :
  1. relâcher les hypothèses sur la loi de la variable aléatoire étudiée qu'on doit faire avec les approches paramétriques
  2. principe générique applicable à n'importe quelle statistique (dont on ne connait pas la distribution)

- Bien garder en tête qu'il y a 2 niveaux d'approximation
  1. remplacer la "vraie" distribution par l'empirique
  2. remplacer la "vraie" distribution d'échantillonnage (selon la loi empirique) par celle obtenue par tirages

- Par conséquent : méthode valide quand n est grand
  - résultats théoriques pour démontrer la validité des procédures décrites (caractérisation et IC)
  - le bootstrap n'est **pas** dédié aux petits échantillons

Remarques / Conclusions

- Intérêts principaux :
  1. relâcher les hypothèses sur la loi de la variable aléatoire étudiée qu'on doit faire avec les approches paramétriques
  2. principe générique applicable à n'importe quelle statistique (dont on ne connait pas la distribution)
- Bien garder en tête qu'il y a 2 niveaux d'approximation
  1. remplacer la "vraie" distribution par l'empirique
  2. remplacer la "vraie" distribution d'échantillonnage (selon la loi empirique) par celle obtenue par tirages
- Par conséquent : méthode valide quand n est grand
  - résultats théoriques pour démontrer la validité des procédures décrites (caractérisation et IC)
  - le bootstrap n'est **pas** dédié aux petits échantillons
- Et si n est petit ? Pas forcément pire qu'une approche paramétrique...

Remarques / Conclusions

- Le principe du bootstrap a été décliné avec succès pour construire des modèles de prédiction, dans une stratégie dite de bagging.

Remarques / Conclusions

- Le principe du bootstrap a été décliné avec succès pour construire des modèles de prédiction, dans une stratégie dite de bagging.
- Un exemple important est celui des forêts aléatoires, qui sont des classifieurs très performants et relativement simples à mettre en oeuvre.

Remarques / Conclusions

- Le principe du bootstrap a été décliné avec succès pour construire des modèles de prédiction, dans une stratégie dite de bagging.
- Un exemple important est celui des forêts aléatoires, qui sont des classifieurs très performants et relativement simples à mettre en oeuvre.
- On peut également utiliser ce principe pour évaluer les performances d'un modèle de prédiction, comme une alternative à la validation croisée.
  - mais c'est moins classique.

⇒ à suivre dans le cours "Apprentissage Statistique II".

Mise en oeuvre R

Le pilier du bootstrap :

```
> ind.bs = sample(n, replace = TRUE)
```

Les packages `bootstrap` et `boot` implémentent les méthodes vues en cours.

Le package `boot` est le plus recommandé.

- extrait tiré de la documentation du package `bootstrap` : *New projects should preferentially use the recommended package "boot"*

Mise en oeuvre R

Package `boot` : deux fonctions principales

1. fonction `boot()`
   - en entrée : données, statistique considérée (une fonction) et B.
   - en sortie : un objet contenant (en 1er lieu) les estimations bootstrap.
   - la fonction `print()` affiche le biais et l'erreur type.
2. fonction `boot.ci()`
   - en entrée : l'objet retourné par la fonction `boot()`.
   - en sortie : les intervalles de confiance par différentes statégies (e.g., `basic`, `perc`, `norm`).

$\Rightarrow$ à voir en TP.

Références

B. Efron and R. Tibshirani. *An introduction to the boostrap*. Chapman & Hall, 1993.