ULB

Une introduction au monde du Bootstrap

Bruffaerts Christopher

European Center for Advanced Research in Economics and Statistics (ECARES)

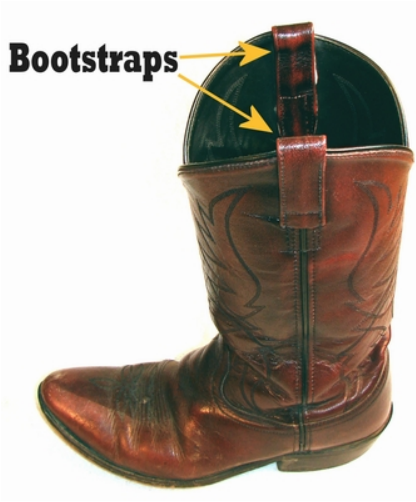
Août 5, 2013

ULB

Quand on google le mot “Bootstrap”...

ULB

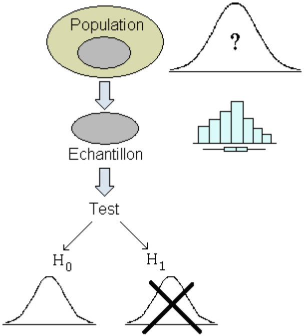
Quand on google le mot “Bootstrap”...



Agenda

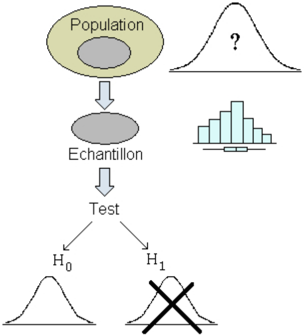
Population VS échantillon

Inférence statistique



Population VS échantillon

Inférence statistique



Domaines d'application de la statistique: Sondages, plans d'expérience, modèles de survie, GLM, contrôle de qualité, statistique spatiale, séries temporelles, . . .

ULB

Paramétrique VS non Paramétrique

Soit un échantillon i.i.d. $\mathcal{X} = (X_1, X_2, \ldots, X_n)$ issu d'une population dont la fonction de répartition est $F(x) = P[X \leq x]$.

ULB

Paramétrique VS non Paramétrique

Soit un échantillon i.i.d. $\mathcal{X} = (X_1, X_2, \ldots, X_n)$ issu d'une population dont la fonction de répartition est $F(x) = P[X \leq x]$.

- $F(.) = F_\theta(.)$ où $\theta \in \mathbb{R}^k \Rightarrow$ paramétrique.
  - La teneur en sucre de bonbons: $X \sim \mathcal{N}(\mu, \sigma^2)$ avec $\theta = (\mu, \sigma^2)'$.
  - Nombre d'accidents sur un laps de temps: $X \sim \mathcal{P}(\lambda)$ avec $\theta = \lambda > 0$.
  - La durée de vie d'ampoules: $X \sim Exp(\lambda)$ avec $\theta = \lambda > 0$.
- $F(.)$ inconnue $\Rightarrow$ non-paramétrique.
  - Tests de signes
  - Tests de rangs

Paramètres d'intérêt

Fonctionnelle statistique: fonction à valeur réelle dont l'argument est une fonction de répartition.

ULB

Paramètres d'intérêt

Fonctionnelle statistique: fonction à valeur réelle dont l'argument est une fonction de répartition.

Exemples:

- *Moyenne*: $\theta = T(F) = \int x dF(x)$.

  ex.: Age moyen des étudiants de l'ULB.

- *Variance*: $\theta = T(F) = \int x^2 dF(x) - \left(\int x dF(x)\right)^2$.

  ex.: Pourcentage de viande de cheval dans les boulettes IKEA.

- *Quantile*: $\theta = T(F) = F^{-1}(\alpha)$, $\alpha \in [0, 1]$.

  ex.: Revenu médian de la population belge.

ULB

Paramètres d'intérêt

Fonctionnelle statistique: fonction à valeur réelle dont l'argument est une fonction de répartition.

Exemples:

- *Moyenne*: $\theta = T(F) = \int x dF(x)$.

  ex.: Age moyen des étudiants de l'ULB.

- *Variance*: $\theta = T(F) = \int x^2 dF(x) - \left(\int x dF(x) \right)^2$.

  ex.: Pourcentage de viande de cheval dans les boulettes IKEA.

- *Quantile*: $\theta = T(F) = F^{-1}(\alpha)$, $\alpha \in [0, 1]$.

  ex.: Revenu médian de la population belge.

Question: Comment estimer ces paramètres à partir de l'échantillon?

F inconnue mais F_n est connue !

ULB

Fonction de répartition empirique

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

Par le théorème de **Glivenko-Cantelli**, $F_n(x) \xrightarrow{P} F(x)$

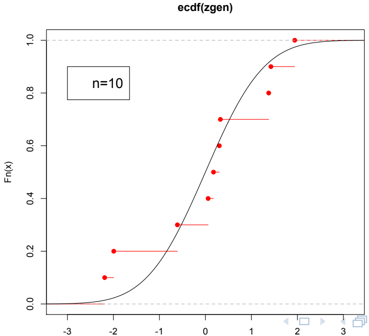
ULB

Fonction de répartition empirique

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

Par le théorème de **Glivenko-Cantelli**, $F_n(x) \xrightarrow{P} F(x)$

Figure: Fonction de répartition de $\mathcal{N}(0,1)$ et sa version empirique.



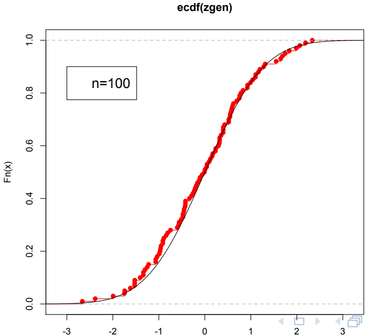
ULB

Fonction de répartition empirique

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

Par le théorème de **Glivenko-Cantelli**, $F_n(x) \xrightarrow{P} F(x)$

Figure: Fonction de répartition de $\mathcal{N}(0,1)$ et sa version empirique.



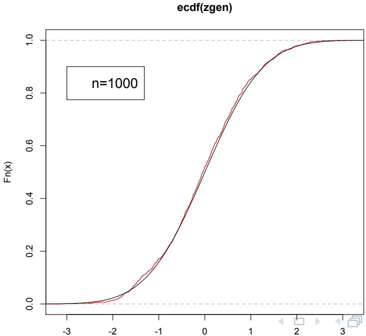
ULB

Fonction de répartition empirique

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

Par le théorème de **Glivenko-Cantelli**, $F_n(x) \xrightarrow{P} F(x)$

Figure: Fonction de répartition de $\mathcal{N}(0,1)$ et sa version empirique.



Estimation: Principe du plug-in

Soit l'échantillon i.i.d. $\mathcal{X} = (X_1, \ldots, X_n)$.

Exemples:

- *Moyenne*: $\widehat{\theta} = T(F_n) = \int x dF_n(x) = \frac{1}{n}\sum_{i=1}^n X_i$.
- *Variance*: $\widehat{\theta} = \int x^2 dF_n(x) - \left(\int x dF_n(x)\right)^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - \left(\frac{1}{n}\sum_{i=1}^n X_i\right)^2$.
- *Médiane*: $\widehat{\theta} = T(F_n) = F_n^{-1}(1/2)$.

ULB

Estimation: Principe du plug-in

Soit l'échantillon i.i.d. $\mathcal{X} = (X_1, \ldots, X_n)$.

Exemples:

- *Moyenne*: $\widehat{\theta} = T(F_n) = \int x dF_n(x) = \frac{1}{n} \sum_{i=1}^n X_i$.
- *Variance*: $\widehat{\theta} = \int x^2 dF_n(x) - \left(\int x dF_n(x) \right)^2 = \frac{1}{n} \sum_{i=1}^n X_i^2 - \left(\frac{1}{n} \sum_{i=1}^n X_i \right)^2$.
- *Médiane*: $\widehat{\theta} = T(F_n) = F_n^{-1}(1/2)$.

Question: Quelle est la distribution échantillonnée de l'estimateur?

$$Biais(\widehat{\theta}) = E_F[\widehat{\theta}] - \theta$$

$$Var(\widehat{\theta}) = E_F\left[\left(\widehat{\theta} - E_F[\widehat{\theta}] \right)^2 \right]$$

$$EQM(\widehat{\theta}) = E_F[(\widehat{\theta} - \theta)^2]$$

ULB

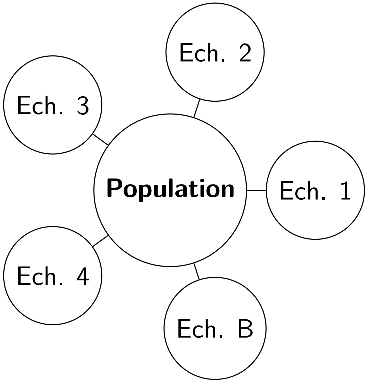
Méthodes de Monte-Carlo

Supposons que F soit **connue**. Soit $S(\mathcal{X})$ une statistique.
Comment obtenir la distribution échantillonnée de S?

Méthodes de Monte-Carlo

Supposons que F soit **connue**. Soit $S(\mathcal{X})$ une statistique.
Comment obtenir la distribution échantillonnée de S?

$\Longrightarrow$ Monte-Carlo: Générer B (grand!) échantillons à partir de F!



ULB

Illustration du Monte-Carlo

Soit $X \sim Exp(\mu)$ et $\theta = \mu = E[X]$, $\quad \widehat{\theta} = S(\mathcal{X}) = \bar{X}$.

1. $\bar{X} \sim \Gamma(n, \frac{\mu}{n})$

$$H_n(x) = P[\bar{X} \leq x] = \int_0^x \frac{u^{n-1} e^{-nu/\mu}}{(\mu/n)^n \Gamma(n)} du$$

2. Pour n grand, $\bar{X} \approx \mathcal{N}(\mu, \mu^2/n)$

$$H_n(x) = P[\bar{X} \leq x] \approx \Phi\left(\frac{x-\mu}{\mu/\sqrt{n}}\right)$$

ULB

Illustration du Monte-Carlo

Soit $X \sim Exp(\mu)$ et $\theta = \mu = E[X]$, $\quad \widehat{\theta} = S(\mathcal{X}) = \bar{X}$.

1. $\bar{X} \sim \Gamma(n, \frac{\mu}{n})$

$$H_n(x) = P[\bar{X} \leq x] = \int_0^x \frac{u^{n-1} e^{-nu/\mu}}{(\mu/n)^n \Gamma(n)} du$$

2. Pour n grand, $\bar{X} \approx \mathcal{N}(\mu, \mu^2/n)$

$$H_n(x) = P[\bar{X} \leq x] \approx \Phi\left(\frac{x - \mu}{\mu/\sqrt{n}}\right)$$

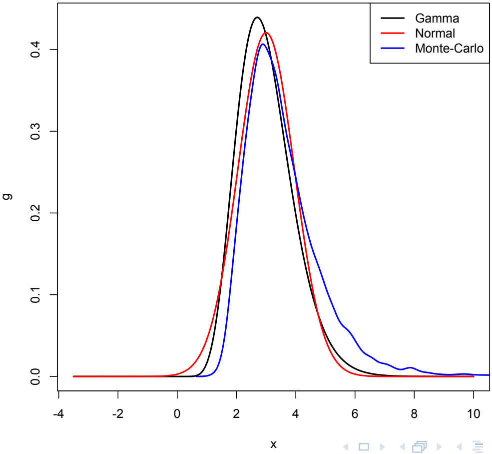
3. Simulations de Monte-Carlo
   - Tirer B échantillons de F (de taille n): $\{\mathcal{X}^{(b)}, b = 1, \ldots, B\}$.
   - Calculer la statistique de test S pour chacun de ces échantillons.
   - Approximation numérique de $H_n(x)$:

$$H_n(x) = P[\bar{X} \leq x] \approx \frac{1}{B} \sum_{b=1}^{B} \mathbb{1}(S(\mathcal{X}^{(b)}) \leq x)$$

ULB

Illustration du Monte-Carlo

Distribution échantillonnée de $\bar{X}$ pour $X \sim Exp(3)$ avec $n = 10$.



Agenda

Deux mondes parallèles

Monde réel

$$\begin{array}{ccc} F & \longrightarrow & F_n \\ T \big\downarrow & & \big\downarrow T \\ \theta & \longleftrightarrow & \widehat{\theta}_n \end{array}$$

$\mathcal{X} = (X_1, X_2, \ldots, X_n)$ issu d'une population dont F est inconnue.

- $T(F_n)$ est un estimateur de θ.
- Distribution de $T(F_n)$ (ou $S(\mathcal{X})$) autour de θ inconnue.

Deux mondes parallèles

Monde réel

$$\begin{array}{ccc} F & \longrightarrow & F_n \\ T\downarrow & & \downarrow T \\ \theta & \longleftrightarrow & \widehat{\theta}_n \end{array}$$

$\mathcal{X} = (X_1, X_2, \ldots, X_n)$ issu d'une population dont F est inconnue.

- $T(F_n)$ est un estimateur de θ.
- Distribution de $T(F_n)$ (ou $S(\mathcal{X})$) autour de θ inconnue.

Monde du bootstrap

$$\begin{array}{ccc} F_n & \rightarrow & F_n^R \\ T\downarrow & & \downarrow T \\ \widehat{\theta}_n & \leftrightarrow & \widehat{\theta}_n^{*R} \end{array}$$

$\mathcal{X}^* = (X_1^*, X_2^*, \ldots, X_n^*)$ issu de la population $\mathcal{X}$ dont F_n est connue.

- $T(F_n^R)$ (ou $S(\mathcal{X}^*)$) est un estimateur de $\widehat{\theta}$.
- **Conditionnellement** à $\mathcal{X}$, la distribution de $S(\mathcal{X}^*)$ est connue!

Deux mondes parallèles

Monde réel

$$\begin{tikzcd} F \arrow[r] \arrow[d, "T"'] & F_n \arrow[d, "T"] \\ \theta \arrow[r, leftrightarrow] & \widehat{\theta}_n \end{tikzcd}$$

$\mathcal{X} = (X_1, X_2, \ldots, X_n)$ issu d'une population dont F est inconnue.

- $T(F_n)$ est un estimateur de θ.
- Distribution de $T(F_n)$ (ou $S(\mathcal{X})$) autour de θ inconnue.

Monde du bootstrap

$$\begin{tikzcd} F_n \arrow[r] \arrow[d, "T"'] & F_n^R \arrow[d, "T"] \\ \widehat{\theta}_n \arrow[r, leftrightarrow] & \widehat{\theta}_n^{*R} \end{tikzcd}$$

$\mathcal{X}^* = (X_1^*, X_2^*, \ldots, X_n^*)$ issu de la population $\mathcal{X}$ dont F_n est connue.

- $T(F_n^R)$ (ou $S(\mathcal{X}^*)$) est un estimateur de $\widehat{\theta}$.
- **Conditionnellement** à $\mathcal{X}$, la distribution de $S(\mathcal{X}^*)$ est connue!

Principe: Distribution de $S(\mathcal{X}) \approx$ Distribution de $S(\mathcal{X}^*)$

Le principe du Bootstrap

L'échantillon bootstrap est à l'échantillon ce que l'échantillon est à la population!

ULB

Le principe du Bootstrap

L'échantillon bootstrap est à l'échantillon ce que l'échantillon est à la population!

Dans le **monde du bootstrap**:

Distribution: $F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(X_i \leq x)$

Paramètre d'intérêt: $\widehat{\theta} = S(\mathcal{X})$.

Echantillon bootstrap $\mathcal{X}^*$ issu de F_n.

Estimateur échantillon bootstrap : $\widehat{\theta}_n^* = S(\mathcal{X}^*)$.

Principe:
Distribution échantillonnée de $\widehat{\theta}_n$ autour de $\theta \approx$ Distribution échantillonnée de $\widehat{\theta}_n^*$ autour de $\widehat{\theta}$.

Le monde du Bootstrap

Question: Comment construire la distribution échantillonnée de $\widehat{\theta}_n^*$?

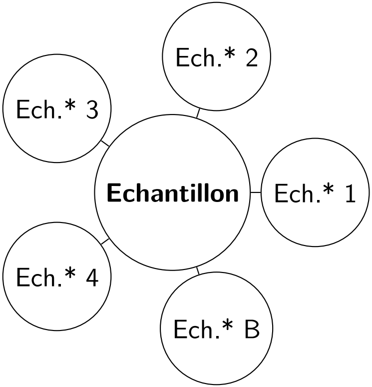
ULB

Le monde du Bootstrap

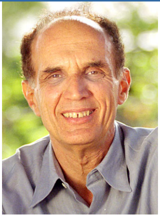
Question: Comment construire la distribution échantillonnée de $\widehat{\theta}_n^*$?

$\Longrightarrow$ via Monte-Carlo (rééchantillonnage):

1. Bootstrap non-paramétrique
2. Bootstrap paramétrique



L'idée du rééchantillonnage (*Efron, 1979*)

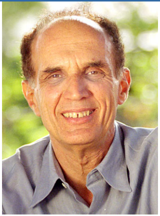


L'idée se base sur la technique de rééchantillonnage avec remplacement.

Qu'est ce qu'un rééchantillonnage?

ULB

L'idée du rééchantillonnage (*Efron, 1979*)



L'idée se base sur la technique de rééchantillonnage avec remplacement.

Qu'est ce qu'un rééchantillonnage?

Soit l'échantillon $\mathcal{X} = (2, 12, 20, 26, 78)$ de taille $n = 5$.

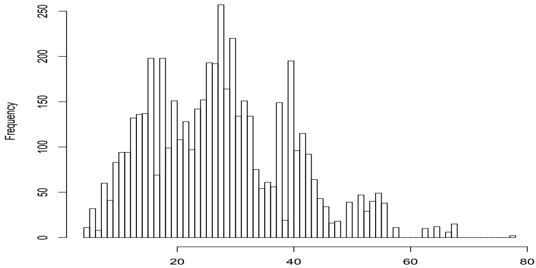
$$\begin{cases} \mathcal{X}^{*(1)} = (12, 20, 78, 26, 2) \\ \mathcal{X}^{*(2)} = (26, 78, 12, 12, 2) \\ \mathcal{X}^{*(3)} = (20, 20, 20, 20, 20) \\ \dots \\ \mathcal{X}^{*(B)} = (26, 2, 2, 78, 20) \end{cases}$$

ULB

Principe du bootstrap: exemple

Soit l'échantillon $\mathcal{X} = (2, 12, 20, 26, 78)$ et $S(\mathcal{X}) = \bar{X}$.

$$\begin{cases} \mathcal{X}^{*(1)} = (12, 20, 78, 26, 2) \longrightarrow S(\mathcal{X}^*) = 27.6 \\ \mathcal{X}^{*(2)} = (26, 78, 12, 12, 2) \longrightarrow S(\mathcal{X}^*) = 26 \\ \mathcal{X}^{*(3)} = (20, 20, 20, 20, 20) \longrightarrow S(\mathcal{X}^*) = 20 \\ \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \\ \mathcal{X}^{*(B)} = (26, 20, 12, 2, 26) \longrightarrow S(\mathcal{X}^*) = 17.2 \end{cases}$$



ULB

Inférence sur le paramètre d'intérêt

Supposons $R_n = \sqrt{n}(\widehat{\theta} - \theta)$ et sa distribution $H_n(x) = P(R_n \leq x)$.

- Si H_n est **connue**,

$$IC(1-\alpha) = \left[\widehat{\theta} - \frac{q_{1-\alpha/2}}{\sqrt{n}}, \widehat{\theta} - \frac{q_{\alpha/2}}{\sqrt{n}}\right]$$

où $q_{\alpha/2} = H_n^{-1}(\alpha/2)$ et $q_{1-\alpha/2} = H_n^{-1}(1-\alpha/2)$.

Exemple: Si $X \sim \mathcal{N}(\mu, \sigma^2)$; $\quad R_n = \sqrt{n}(\bar{X} - \mu) \sim \mathcal{N}(0, \sigma^2)$

$$IC(1-\alpha) = \left[\bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

Inférence sur le paramètre d'intérêt

Supposons $R_n = \sqrt{n}(\widehat{\theta} - \theta)$ et sa distribution $H_n(x) = P(R_n \leq x)$.

- Si H_n est **connue**,

$$IC(1-\alpha) = \left[\widehat{\theta} - \frac{q_{1-\alpha/2}}{\sqrt{n}}, \widehat{\theta} - \frac{q_{\alpha/2}}{\sqrt{n}}\right]$$

où $q_{\alpha/2} = H_n^{-1}(\alpha/2)$ et $q_{1-\alpha/2} = H_n^{-1}(1-\alpha/2)$.

Exemple: Si $X \sim \mathcal{N}(\mu, \sigma^2)$; $\quad R_n = \sqrt{n}(\bar{X} - \mu) \sim \mathcal{N}(0, \sigma^2)$

$$IC(1-\alpha) = \left[\bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

- Si H_n est **inconnue**, le bootstrap approxime H_n!
Quantiles $q^*_{\alpha/2} = \widehat{H}_n^{-1}(\alpha/2)$ et $q^*_{1-\alpha/2} = \widehat{H}_n^{-1}(1-\alpha/2)$ où:

$$\widehat{H}_n(x) \approx \frac{1}{B}\sum_{b=1}^{B} \mathbb{1}\left(\sqrt{n}(\widehat{\theta}^{*(b)} - \widehat{\theta}) \leq x\right)$$

ULB

Un peu d'histoire...

Baron Münchausen: militaire allemand (1720-1797).

Origine: Provient des histoires à dormir debout du Baron Münchausen. Il prétend que son cheval et lui se soient sortis d'un marécage simplement en tirant sur:

ULB

Un peu d'histoire...

Baron Münchausen: militaire allemand (1720-1797).

Origine: Provient des histoires à dormir debout du Baron Münchausen. Il prétend que son cheval et lui se soient sortis d'un marécage simplement en tirant sur:

1. ses cheveux
2. les lanières de ses bottes (bootstrap)

"to pull oneself up by one's own bootstraps"

$\Rightarrow$ Analogie avec le problème statistique:

on se débrouille **uniquement** avec l'échantillon!

ULB

Un peu d'histoire...

Figure: Baron Münchausen (1720-1797)

Agenda

ULB

De quoi a-t-on besoin pour “bootstrapper”?

“Dans le bootstrap, on a besoin de rien...”

...sauf peut-être...

ULB

De quoi a-t-on besoin pour “bootstrapper”?

“Dans le bootstrap, on a besoin de rien...”

...sauf peut-être...

Sources de variations dans le bootstrap:

1. échantillon original aléatoire
   $\Rightarrow$ Approximation de $H_n(x)$ par $\widehat{H}_n(x)$.
2. tirage aléatoire des échantillons bootstrap
   $\Rightarrow$ Approximation de $\widehat{H}_n(x)$ via Monte-Carlo (contrôle sur B!)

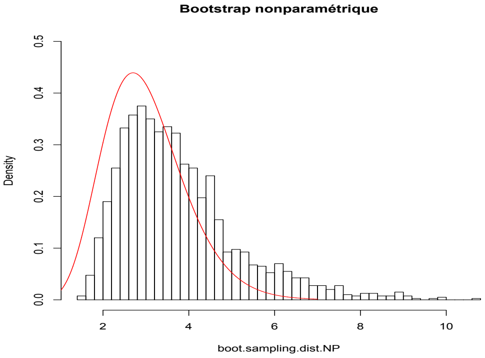
Bootstrap non paramétrique

$$
\begin{tikzcd}
\text{Echantillon de départ } \mathcal{X} = (X_1, \ldots, X_n) \arrow[r, dashed, "\text{Rééchantillonnage}"] \arrow[d] & \text{Echantillon Bootstrap } \mathcal{X}^* = (X_1^*, \ldots, X_n^*) \arrow[d] \\
\text{Distribution empirique } F_n(.) \arrow[d, "S"] & \text{Distribution } P[X_i^* \leq x \mid \mathcal{X}] = F_n(x) \arrow[d, "S"] \\
\widehat{\theta} = S(\mathcal{X}) & \widehat{\theta}^* = S(\mathcal{X}^*)
\end{tikzcd}
$$

ULB

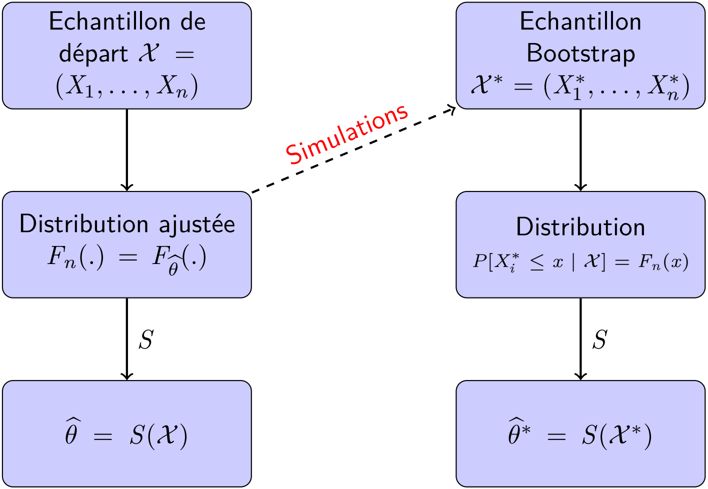
Bootstrap non paramétrique: exemple

Soit l'échantillon $(1.7, 2, 2.4, 3.9, 4.3, 4.5, 5.2, 5.8, 6.2, 7.4)$; $n = 10$.
Soit $S(\mathcal{X}) = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.



Distribution exacte: $\bar{X} \sim \Gamma(n, \frac{\mu}{n})$; $\quad (X \sim Exp(\mu))$.

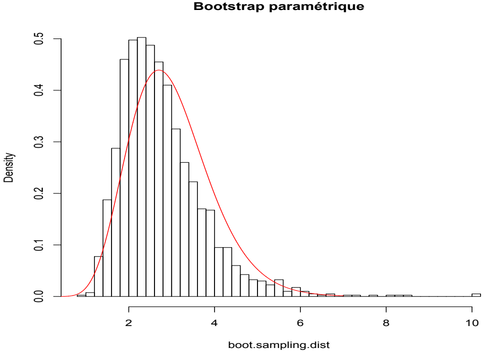
Bootstrap paramétrique



ULB

Bootstrap paramétrique: exemple

Soit l'échantillon $(1.2, 1.7, 1.8, 2, 2.3, 2.6, 3.9, 4.3, 4.5, 5.8)$; $n = 10$.
Simuler $B = 5000$ échantillons de la distribution $Exp(\bar{x})$.



Distribution exacte: $\bar{X} \sim \Gamma(n, \frac{\mu}{n})$; $\quad (X \sim Exp(\mu))$.

ULB

Algorithme du bootstrap

Soit l'échantillon $\mathcal{X} = (X_1, \ldots, X_n)$.

1. Générer un échantillon bootstrap $\mathcal{X}^* = (X_1^*, \ldots, X_n^*)$:
   - Paramétrique
   - Non-paramétrique
2. Calculer $S^* = S(X_1^*, \ldots, X_n^*) = \widehat{\theta}^*$.
3. Répéter les étapes 1 et 2 un certain nombre de fois B.
4. Approximer $H_n(x)$ par

$$\widehat{H}_n(x) = \frac{1}{B} \sum_{j=1}^{B} \mathbb{1}\Big(\sqrt{n}(\widehat{\theta}_j^* - \widehat{\theta}) \leq x\Big).$$

5. Trouver les quantiles $\widehat{q}_{\alpha/2}$ et $\widehat{q}_{1-\alpha/2}$ de $\widehat{H}_n(x)$ et construire l'IC.

Agenda

Le bootstrap: une panacée pour l'inférence?

Forces:

- Permet de calculer des intervalles de confiance, de corriger pour le biais et d'estimer la variabilité liée à l'estimation.
- Peut être une solution pour des problèmes compliqués.
- Même si les résultats asymptotiques sont disponibles, le bootstrap offre souvent de meilleures approximations.
- Facile à implémenter.
- Flexible.

Faiblesses:

- Ne marche pas toujours!
- Peut être lourd computationnellement.
- Dépend de la représentativité de l'échantillon de départ.

ULB

Agenda

ULB

Exemple 1: Régression linéaire (1)

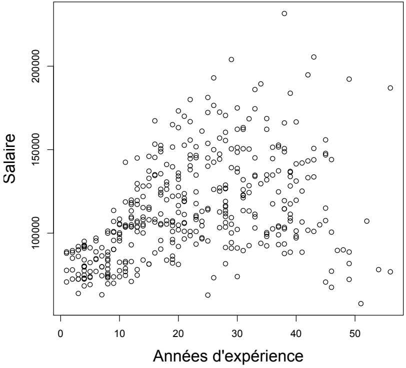
Modèle: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, n; \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma^2).$

ULB

Exemple 1: Régression linéaire (1)

Modèle: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, n; \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma^2).$

Données: Salaires de 397 professeurs dans un "American College" (2008-2009).

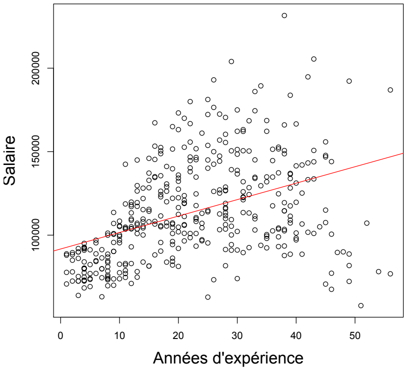


ULB

Exemple 1: Régression linéaire (1)

Modèle: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, n; \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma^2).$

Données: Salaires de 397 professeurs dans un "American College" (2008-2009).



ULB

Exemple 1: Régression linéaire (2)

Inférence: $$\frac{\widehat{\beta}_0 - \beta_0}{SE(\widehat{\beta}_0)} \sim t_{n-1} \quad \text{et} \quad \frac{\widehat{\beta}_1 - \beta_1}{SE(\widehat{\beta}_1)} \sim t_{n-1}.$$

ULB

Exemple 1: Régression linéaire (2)

Inférence: $$\frac{\widehat{\beta_0}-\beta_0}{SE(\widehat{\beta_0})} \sim t_{n-1} \quad \text{et} \quad \frac{\widehat{\beta_1}-\beta_1}{SE(\widehat{\beta_1})} \sim t_{n-1}.$$

Possibilités de rééchantillonnage:

1. couples $\{(x_i, y_i), i = 1, \ldots, n\}$
2. résidus $\{r_i = y_i - \widehat{y}_i, i = 1, \ldots, n\}$

	β_0		β_1	
	Est.	*SE*	*Est.*	*SE*
Théorie	91718.7	2765.8	985.3	107.4
Bootstrap	91718.6	2454.1	985.3	126

	β_0 $IC(95\%)$	β_1 $IC(95\%)$
Théorie	[86281, 97156]	[774, 1196]
Bootstrap	[87315, 97726]	[702, 1225]

Exemple 2: estimation d'une frontière (1)

Contexte: $X \sim U[0, \theta]$ et $\mathcal{X} = (X_1, X_2, \ldots, X_n)$.

$$\widehat{\theta}_{MV} = \max(X_1, X_2, \ldots, X_n)$$

ULB

Exemple 2: estimation d'une frontière (1)

Contexte: $X \sim U[0, \theta]$ et $\mathcal{X} = (X_1, X_2, \ldots, X_n)$.

$$\widehat{\theta}_{MV} = \max(X_1, X_2, \ldots, X_n)$$

Monde réel:

$$R_n(\theta) = \frac{n(\theta - \widehat{\theta})}{\theta} \sim Exp(.)$$

Monde du bootstrap:

$$R_n^*(\widehat{\theta}) = \frac{n(\widehat{\theta} - \widehat{\theta}^*)}{\widehat{\theta}}$$

$$P[R_n^*(\widehat{\theta}) = 0 \mid \mathcal{X}] = 1 - \left(1 - \tfrac{1}{n}\right)^n \longrightarrow 1 - e^{-1} \quad \text{pour } n \to \infty.$$

Problème: le maximum est tiré trop souvent dans le monde du bootstrap.

ULB

Exemple 2: estimation d'une frontière (2)

Contexte: $X \sim U[0, \theta]$ et $\mathcal{X} = (X_1, X_2, \ldots, X_n)$.

Solution: sous-échantillonnage ("subsampling")

Monde réel:

$$R_n(\theta) = \frac{n(\theta - \widehat{\theta})}{\theta} \sim Exp(.)$$

Monde du bootstrap:

$$R_n^*(\widehat{\theta}) = \frac{n(\widehat{\theta} - \widehat{\theta}^*)}{\widehat{\theta}}$$

$P[R_n^*(\widehat{\theta}) = 0 \mid \mathcal{X}] = 1 - \left(1 - \frac{1}{n}\right)^m \longrightarrow 0$ pour $n \to \infty$, et $\frac{m}{n} \to 0$, et où m est la taille du sous-échantillon.

ULB

Agenda

ULB

Conclusion

- Le bootstrap est une technique prometteuse fortement utilisée (l'usage computationel s'améliore).
- ⚠à ne pas banaliser l'usage du bootstrap.
- Le bootstrap en pratique:
  1. **R cran**: package *boot* et *bootstrap*
  2. **Matlab**: bootstrap toolbox
- Extensions possibles:
  - Subsampling
  - Smooth bootstrap
  - Block bootstrap (séries temporelles)
  - Jackknife
  - Iterated bootstrap
  - ...

ULB

La conclusion de John Tukey

"The bootstrap is like a **shotgun** because you can blow the head off any statistical problem..."

ULB

Quelques références

- P. Bickel and D. Freeman (1981) *Some asymptotic theory for the bootstrap*, *Annals of Statistics*, 9:1196-1217.
- A.C. Davison and D.V. Hinkley (1997) *Bootstrap methods and their Application*, *Cambridge University Press*.
- B. Efron (1979) *Bootstrap methods: Another look at the jackknife*, *Annals of Statistics*, 7:1-26.
- B. Efron and R.J. Tibshirani (1993) *An introduction to the bootstrap*, Chapman and Hall, New York.
- P. Hall (1992) *The bootstrap and Edgeworth Expansion*, Springer-Verlag, New York.
- J.L. Horowitz (2000) *The bootstrap*.
- L. Simar (2008) *An invitation to the bootstrap: Panacea for Statistical Inference?*