Bootstrap et rééchantillonnage

Atelier SFdS - Partie 1

Magalie Fromont et Myriam Vimond

CREST (Ensai) - IRMAR (Université Européenne de Bretagne)

Novembre 2012

Plan

Il était une fois...

Bradley Efron

Bootstrap methods : another look at the Jackknife. Ann. Stat. (1979)

Question

Etant donné un échantillon $\mathbb{X} = (X_1, \ldots, X_n)$ de variables aléatoires i.i.d. de loi P inconnue, comment estimer la loi d'une variable $R_n(\mathbb{X}, P)$ sur la base d'une observation de $\mathbb{X}$?

S'inspirant du Jackknife, introduit par Quenouille et Tukey pour l'estimation du biais et de la variance d'un estimateur, Efron propose une méthode d'estimation non paramétrique plus "primitive" (sic), mais plus générale.

Il était une fois...

Karl Friedrich Hieronymus, baron de Münchhausen ?

Le terme de bootstrap donné par Efron à sa méthode provient de l'expression : "to pull oneself up by one's bootstraps".

↪ Se hisser en tirant sur les languettes de ses bottes.
↪ Se sortir, seul, d'une situation difficile.

Cette expression est communément attribuée à Rudolf E. Raspe, auteur des aventures du baron de Münchhausen (*Baron Münchhausen's Narrative of his Marvellous Travels and Campaigns in Russia*, 1785).

Dans une lontaine contrée...

Du doux nom d'Inférence Statistique

Rappel : $\mathbb{X} = (X_1, \ldots, X_n)$ échantillon d'une loi P inconnue.

On s'intéresse à un paramètre $\theta(P)$ de la loi P et on envisage l'un des problèmes suivants.

- Estimation ponctuelle de $\theta(P)$ par $\hat{\theta} = T(\mathbb{X})$ et étude de la précision de l'estimateur :
  - Biais de $\hat{\theta} = \mathbb{E}_P[T(\mathbb{X}) - \theta(P)]$,
  - Variance de $\hat{\theta} = \mathbb{V}_P[T(\mathbb{X})]$,
  - EQM (MSE) de $\hat{\theta} = \mathbb{E}_P[(T(\mathbb{X}) - \theta(P))^2]$.
- Construction d'une région de confiance pour $\theta(P)$.
- Construction d'un test d'hypothèses sur $\theta(P)$.

Pour résoudre le problème envisagé, il "suffit" par exemple de :

- Connaître (exactement ou asymptotiquement) la loi de $T(\mathbb{X}) - \theta(P)$, de $T(\mathbb{X})$ ou de $\phi\left((T(\mathbb{X}) - \theta(P)\right)$.

OU

- Savoir simuler cette loi de façon à pouvoir utiliser des méthodes de Monte Carlo.

L'exemple de la moyenne.

$\mathbb{X} = (X_1, \ldots, X_{30})$ 30-échantillon d'une loi P.

Paramètre d'intérêt : $\theta(P) = \mathbb{E}_P[X_i] = \int x dP(x)$.

Estimateur des moments, du maximum de vraisemblance et plug-in : $\hat{\theta} = \overline{\mathbb{X}}$.

$\hookrightarrow$ Question : loi de $\hat{\theta} = \overline{\mathbb{X}}$?

Loi P = loi de Bernoulli $\mathcal{B}(1/2)$ (connue).

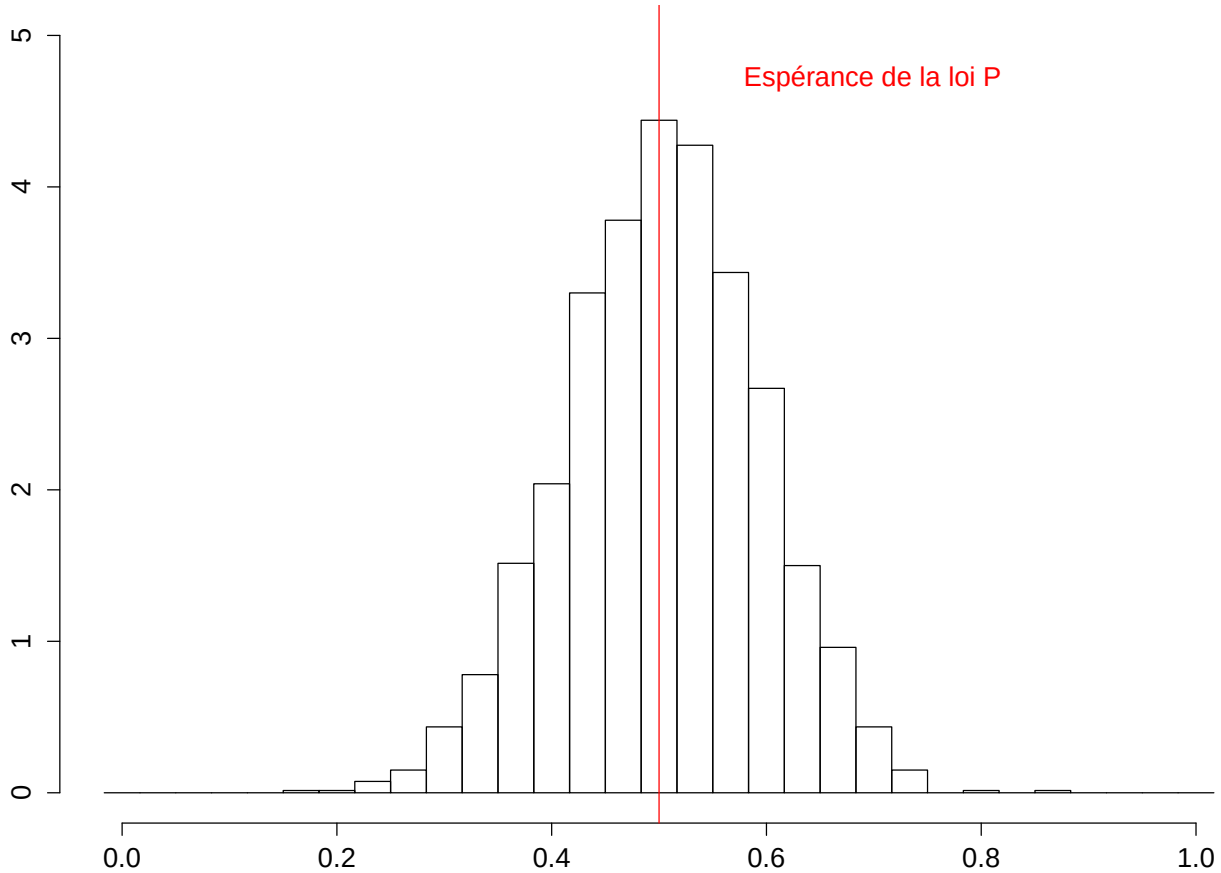

Figure: Histogramme de la loi de $\hat{\theta} = \overline{\mathbb{X}}$

Loi P = loi de Bernoulli $\mathcal{B}(\theta)$ (θ inconnu).
$\widehat{\mathbb{X}}$ = 30-échantillon de loi $\mathcal{B}(\hat{\theta})$ conditionnellement à $\mathbb{X}$.

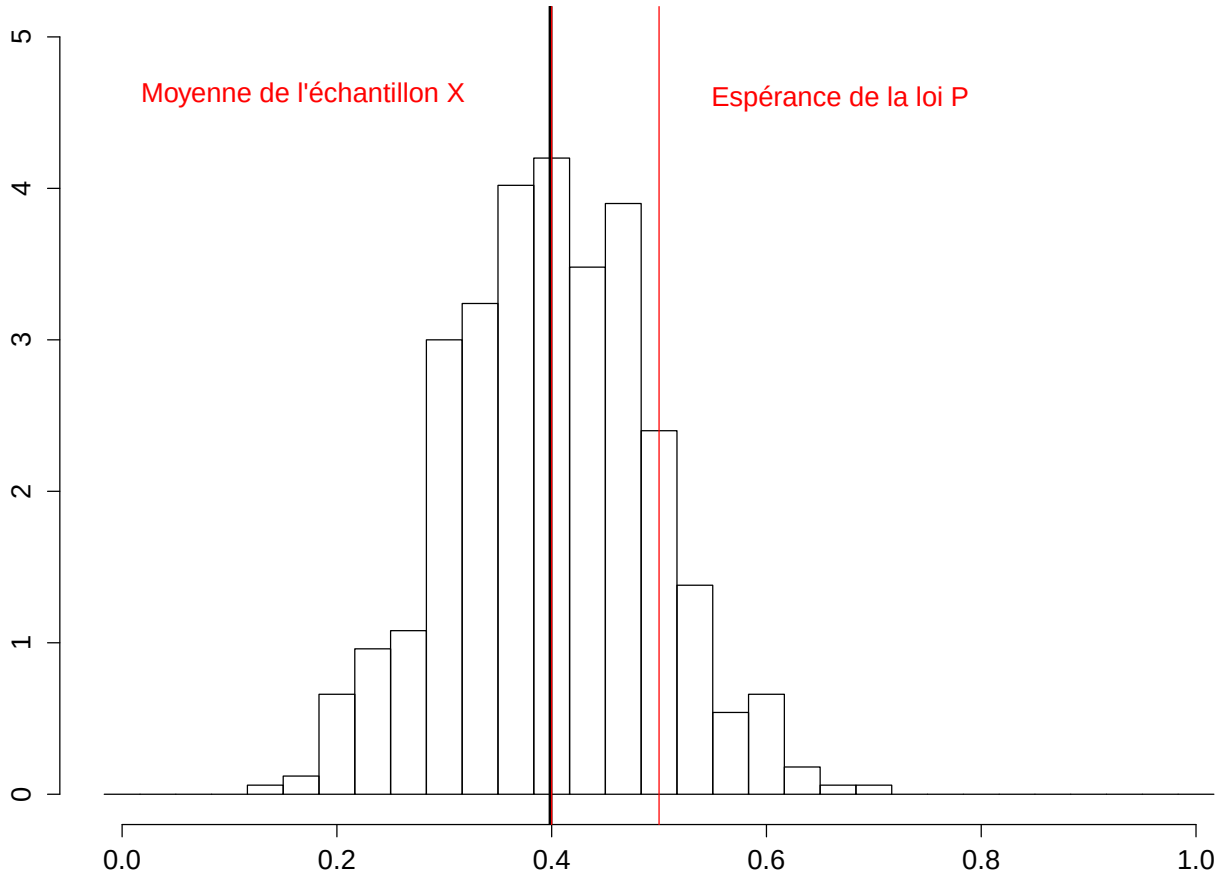

Figure: Histogramme de la loi de $\overline{\widehat{\mathbb{X}}}$ sachant $\mathbb{X}$

Loi P inconnue.
$\mathbb{X}^* = (X_1^*, \ldots, X_{30}^*)$, X_i^* tiré au hasard avec remise dans $\{X_1, \ldots, X_{30}\}$.

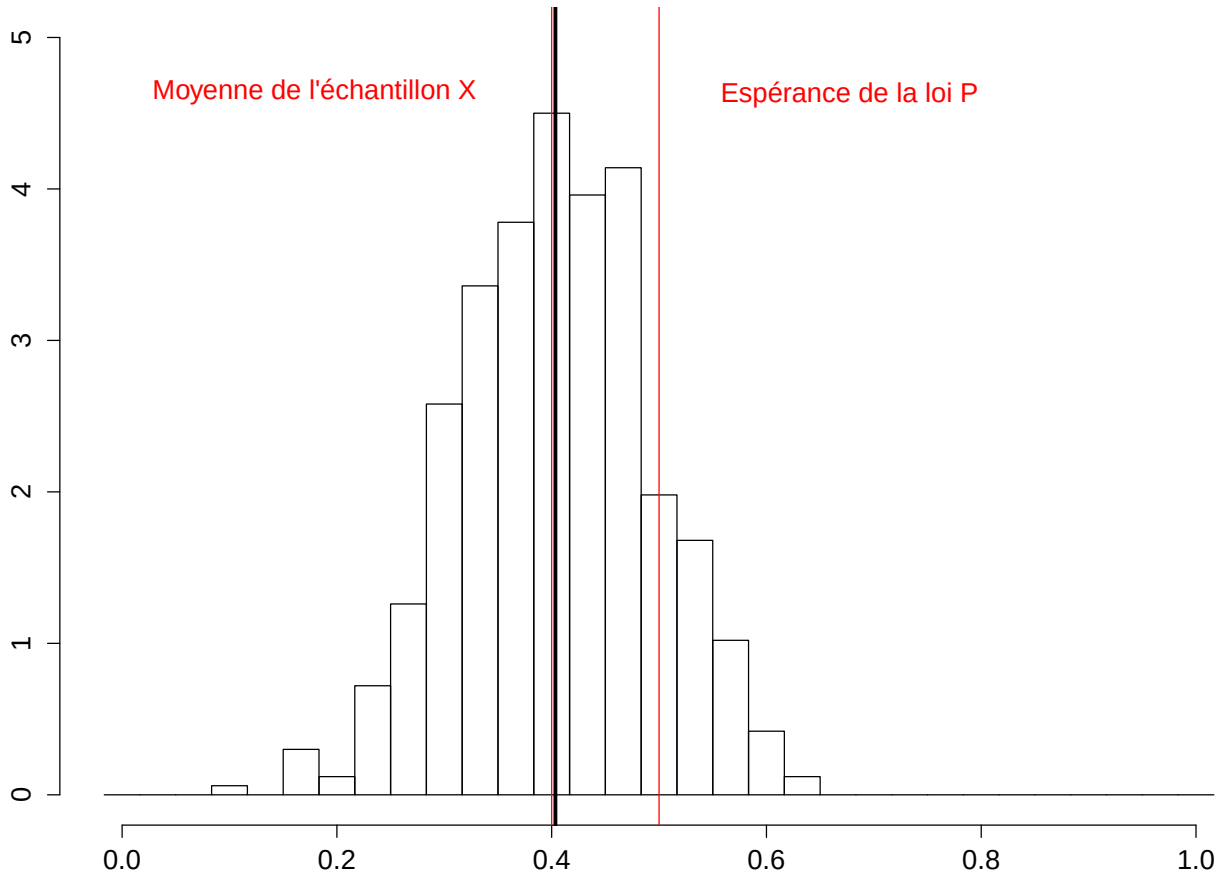

Figure: Histogramme de la loi de $\overline{\mathbb{X}^*}$ sachant $\mathbb{X}$

Loi $P =$ loi de Bernoulli $\mathcal{B}(1/2)$ (connue).
$R(\mathbb{X}, P) = \sqrt{30}\left(\overline{\mathbb{X}} - \theta(P)\right)$.

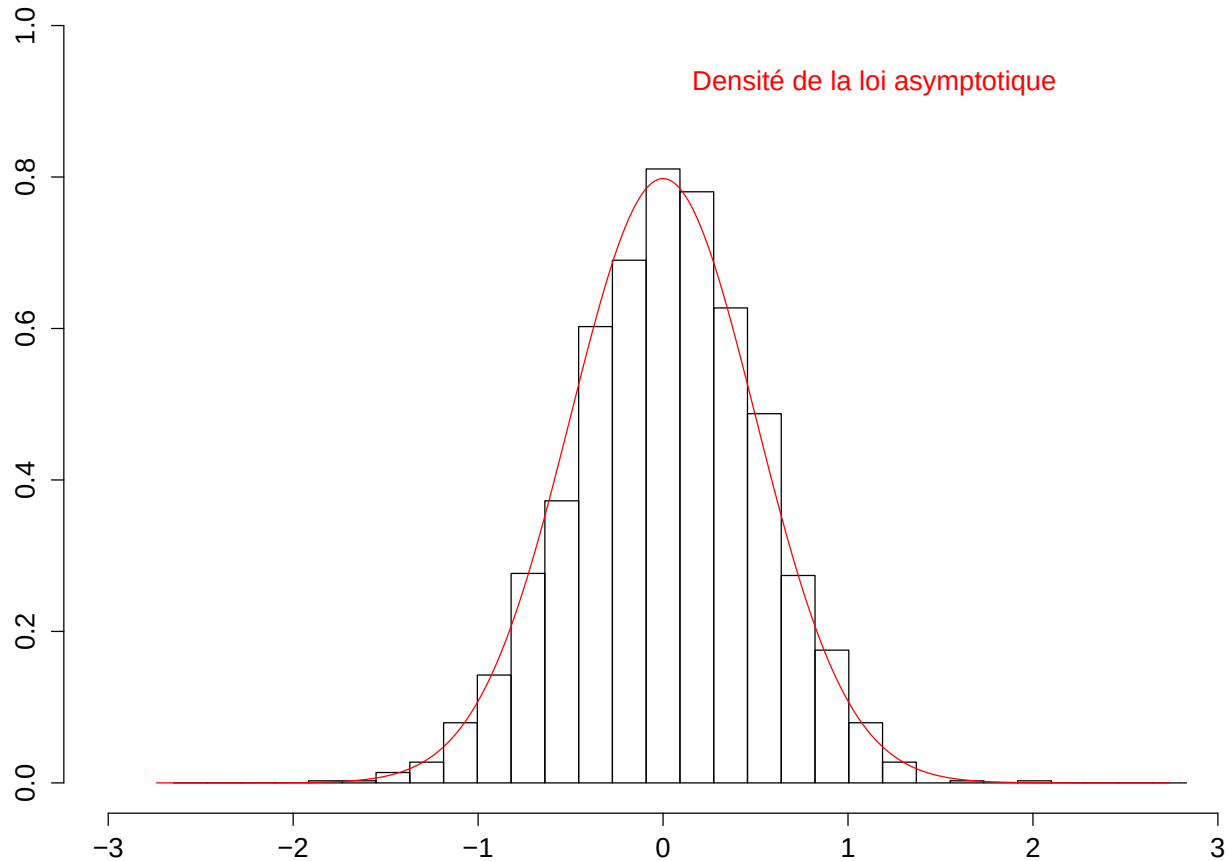

Figure: Histogramme de la loi de $R(\mathbb{X}, P)$

Loi $P =$ loi de Bernoulli $\mathcal{B}(\theta)$ (θ inconnu).

$\widehat{R} = \sqrt{30}\left(\overline{\widehat{\mathbb{X}}} - \hat{\theta}\right)$.

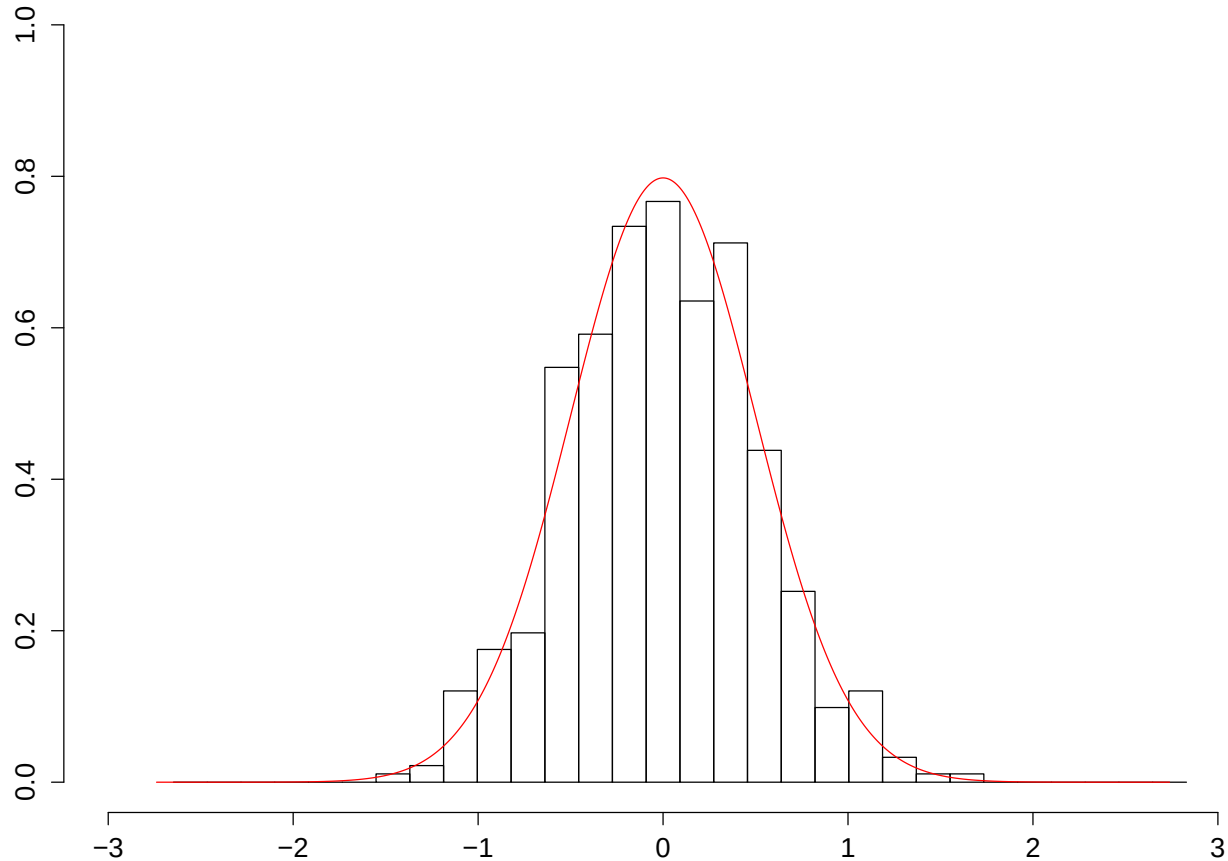

Figure: Histogramme de la loi de $\widehat{R}$ sachant $\mathbb{X}$

Loi P inconnue.
$R^* = \sqrt{30}\left(\overline{\mathbb{X}^*} - \overline{\mathbb{X}}\right)$.

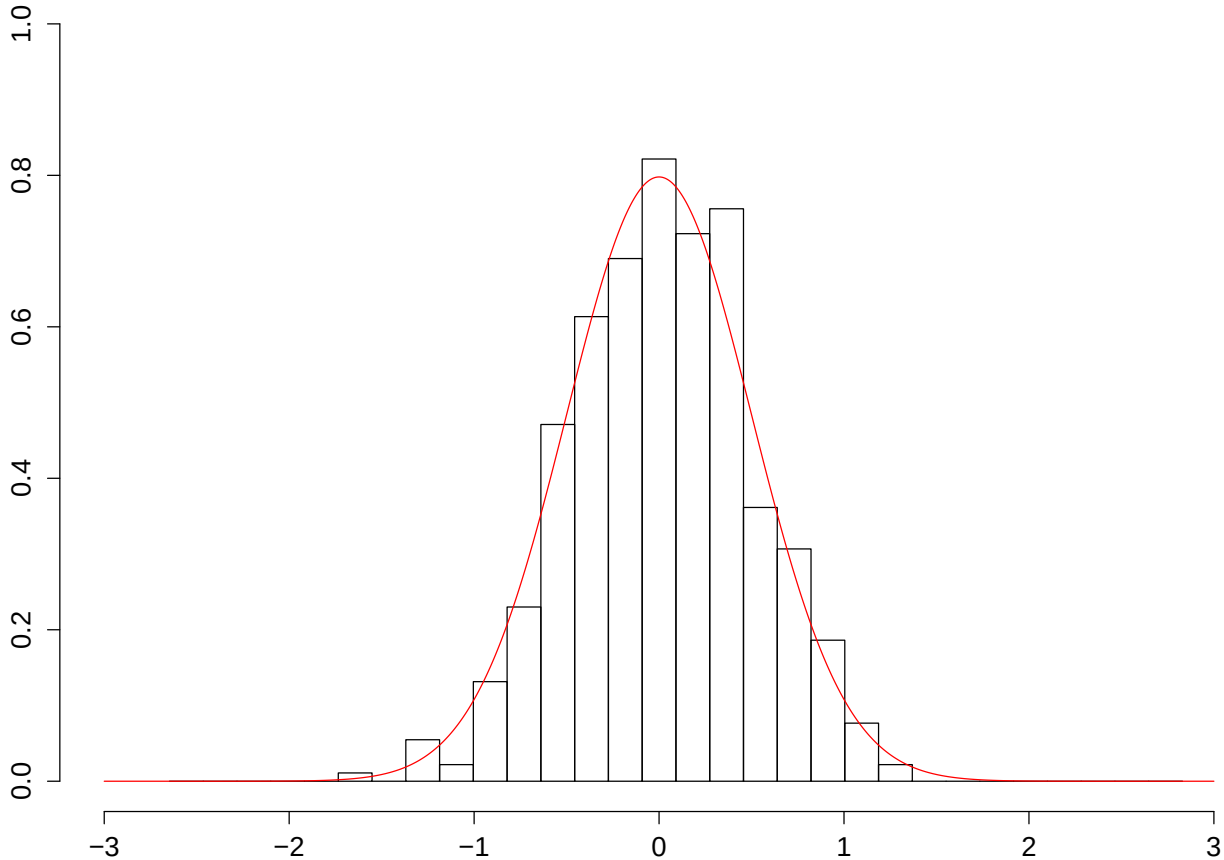

Figure: Histogramme de la loi de R^* sachant $\mathbb{X}$

Ils vécurent heureux...

Quelques ouvrages de référence

- Chernick, M. R. (1999). Bootstrap Methods : A Practitioner's Guide.
- Davison, A. C. and Hinkley, D. V. (1997). Bootstrap Methods and their Applications.
- Efron, B. and Tibshirani, R. J. (1993). An introduction to the Bootstrap.
- Giné, E. (1996). Lectures on some aspects of the Bootstrap.
- Hall, P. (1992). The Bootstrap and Edgeworth Expansion.
- Shao, J., and Tu, D. (1992). The Jackknife and Bootstrap.

Plan

La mesure empirique

La fonction de répartition empirique

Soit $(\Omega, \mathcal{A}, \mathbb{P})$ un espace probabilisé.
Soit $\mathbb{X} = (X_1, \ldots, X_n)$ un n-échantillon de P.

Fonction de répartition : $F(t) = \mathbb{P}(X_1 \leq t)$

Fonction de répartition empirique : $F_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{X_i \leq t}$

La **mesure empirique** P_n est la mesure discrète associée à F_n qui attribue un poids de $1/n$ à chaque observation :

$$P_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i},$$

où δ_x est la mesure de Dirac en x.

La mesure empirique

La mesure empirique comme un processus

Notations : soit $g : \mathbb{R} \to \mathbb{R}$ fonction mesurable bornée,

$$Pg = \int g dP = \mathbf{E}[g(X_1)], \qquad P_n g = \int g dP_n = \frac{1}{n} \sum_{i=1}^{n} g(X_i).$$

Par exemple : $F(t) = P\mathbb{1}_{]-\infty,t]}$, $F_n(t) = P_n\mathbb{1}_{]-\infty,t]}$.

Soit $\mathcal{F} \subset L^2(P)$.
La **mesure empirique** indexée par $\mathcal{F}$, $P_n = (P_n f)_{f\in\mathcal{F}}$, est une collection de v.a. indexée par $\mathcal{F}$.

Convergence ponctuelle du processus

(LFG) Pour tout $f \in \mathcal{F}$, $P_n f \xrightarrow{p.s.} Pf$.

Hypothèse : Les trajectoires de P_n sont p.s. bornées

Pour presque tout $\omega \in \Omega$ fixé, l'application

$$\begin{aligned} \mathcal{F} \subset L^2(P) &\to \mathbb{R} \\ f \quad &\mapsto (P_n f)(\omega) = \frac{1}{n} \sum_{i=1}^{n} f(X_i(\omega)) \end{aligned}$$

est bornée.

Objectif : Convergence sur les trajectoires de P_n.

$$\|P_n - P\|_\infty = \sup_{f \in \mathcal{F}} \left| P_n f - P f \right| \text{ mesurable ? convergente ?}$$

Exemple : $\mathcal{F} = \left\{ \mathbb{1}_{]-\infty,t]}, t \in \mathbb{R} \right\}$

Théorème de Glivenko-Cantelli (1933)

$$\|P_n - P\|_\infty = \sup_t |F_n(t) - F(t)| \overset{p.s.}{\to} 0$$

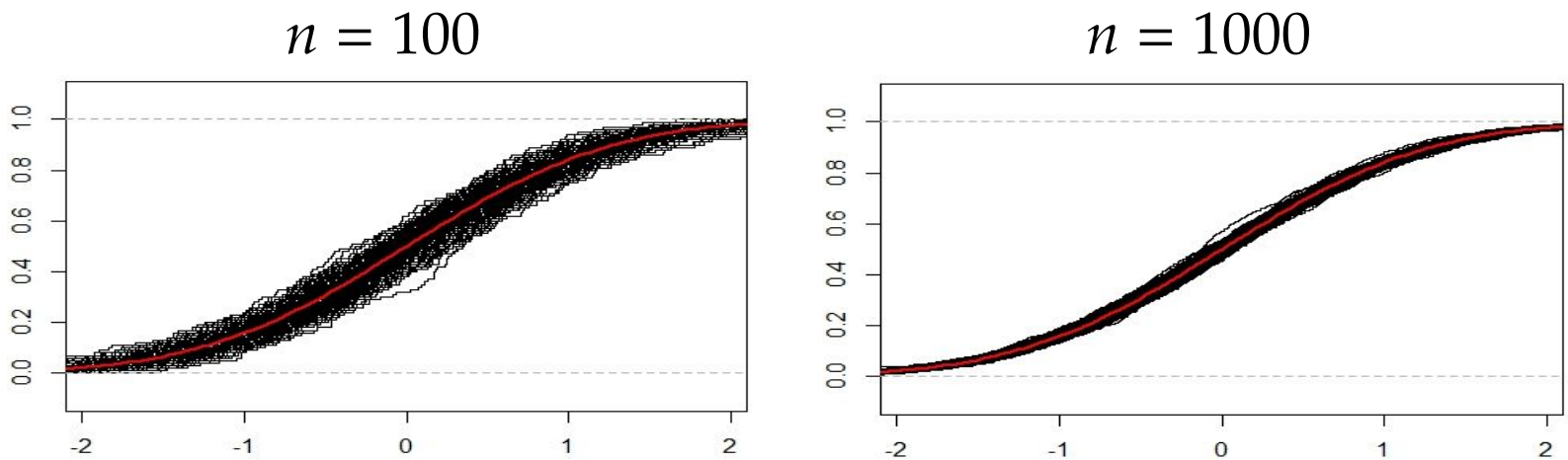

Figure: 200 trajectoires de la mesure empirique P_n associée à un n-échantillon de la loi $\mathcal{N}(0,1)$

Le processus empirique

Le **processus empirique** indexé par $\mathcal{F}$, $Z_n = (Z_n f)_{f\in\mathcal{F}}$, est une collection de v.a. indexée par $\mathcal{F}$, telle que :

$$\text{pour } f \in \mathcal{F}, \quad Z_n f = \sqrt{n}\,(P_n f - P f)\,.$$

Convergence ponctuelle du processus empirique

(TCL) pour tout $f \in \mathcal{F} \subset L^2(P)$, $\quad Z_n f \rightsquigarrow Z f$

où $Z f \sim \mathcal{N}\left(0, Pf^2 - (Pf)^2\right)$.

Objectif : Convergence sur les trajectoires de Z_n lorsque ces trajectoires sont p.s. bornées

$$Z_n : \Omega \to \ell^\infty(\mathcal{F}),$$

où $\ell^\infty(\mathcal{F}) = \left\{z : \mathcal{F} \to \mathbb{R} :\ \|z\|_\infty = \sup_f |z(f)| < \infty\right\}$.

Exemple : $\mathcal{F} = \left\{\mathbb{1}_{]-\infty,t]}, t \in \mathbb{R}\right\}$

Inégalité de Dvoretsky-Kiefer-Wolfowitz

(**Bande de confiance pour** $(F(t))_{t\in\mathbb{R}}$) Pour tout $\epsilon > 0$,

$$\lim_n \tilde{\mathbb{P}}\left(\|Z_n\|_\infty > \epsilon\right) = \lim_n \tilde{\mathbb{P}}\left(\|F_n - F\|_\infty > \epsilon/\sqrt{n}\right) \leq 2e^{-2\epsilon^2}.$$

Théorème de Donsker (1952)

$$Z_n = \sqrt{n}(P_n - P) \rightsquigarrow Z \quad \text{dans } \ell^\infty(\mathbb{R}),$$

Z processus gaussien centré de covariance $\sigma(s,t) = F(s\wedge t) - F(s)F(t)$.

Si X est un n-échantillon de la loi $\mathcal{U}[0,1]$,

Z est un pont brownien sur $[0,1]$.

Définitions :

Soit $Y_n : \Omega \to \ell^\infty(\mathcal{F})$ une suite d'applications.
Soit $Y : \Omega \to \ell^\infty(\mathcal{F})$ une v.a.

- $\mathbf{Y_n \xrightarrow{p.s.} Y}$ dans $\ell^\infty(\mathcal{F})$
  s'il existe $\Delta_n : \Omega \to \mathbb{R}_+$ v.a. telle que $\Delta_n \xrightarrow{p.s.} 0$ et $\|Y_n - Y\|_\infty \leq \Delta_n$.

- $\mathbf{Y_n \xrightarrow{\mathbb{P}} Y}$ dans $\ell^\infty(\mathcal{F})$
  si pour tout $\epsilon > 0$, $\tilde{\mathbb{P}}\left(\|Y_n - Y\|_\infty > \epsilon\right) \to 0$.

- $\mathbf{Y_n \rightsquigarrow Y}$ dans $\ell^\infty(\mathcal{F})$
  si pour toute $f : \ell^\infty(\mathcal{F}) \to \mathbb{R}$ continue bornée,
  $\tilde{\mathbb{E}}\left(f(Y_n)\right) \to \mathbb{E}\left(f(Y)\right)$.

Méthode d'estimation par injection : *plug-in*

Fonctionnelle statistique

Une **fonctionnelle statistique** est un paramètre s'exprimant comme une fonction de la loi de l'échantillon

$$\theta = \theta(P) \quad \text{ou} \quad \theta = \theta(F).$$

Exemples :

- La moyenne $m(P) = \int x dP$
- La variance $\sigma^2(P) = \int (x - m(P))^2 \, dP$
- La médiane $m_e(P) = F^{-1}(1/2)$
- Le coefficient d'asymétrie $\gamma(P) = \int (x - m(P))^3 \, dP / \sigma(P)^{3/2}$
- Fonctionnelle linéaire statistique : pour $f \in L^1(P)$ fixée,

$$s(P) = \int f(x) dP(x)$$

Méthode d'estimation par injection : *plug-in*

Un **estimateur par injection** (ou **plug-in**) est défini par :

$$\hat{\theta} = \theta(P_n)$$

Exemples :

- La moyenne $m(P_n) = \overline{X}$
- La variance $\sigma^2(P_n) = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$
- La médiane $m_e(P_n) = X_{(n/2)}$ ou $X_{((n+1)/2)}$ i.e. $X_{(\lceil n/2 \rceil)}$

  $\vdots$
- Fonctionnelle linéaire statistique : $s(P_n) = P_n f$

Zoom sur : $\theta(P) = Pf$ avec $f \in L^2(P)$.

- Estimateur plug-in de $\theta(P)$: $\hat{\theta} = \theta(P_n) = P_n f$
  LFG : $\hat{\theta} \overset{p.s.}{\to} \theta(P)$
  TCL : $\sqrt{n}(\hat{\theta} - \theta) \rightsquigarrow \mathcal{N}(0, \sigma^2)$, avec $\sigma^2 = Pf^2 - (Pf)^2$

- Estimateur plug-in de σ^2 : $\hat{\sigma}^2 = P_n f^2 - (P_n f)^2$
  LFG : $\hat{\sigma}^2 \overset{p.s.}{\to} \sigma^2$

Soit $z_{\alpha/2}$ le quantile d'ordre $\alpha/2$ de la loi $\mathcal{N}(0,1)$.
Un intervalle de confiance de niveau de confiance asymptotique $1-\alpha$ pour $\theta(P)$ est donné par :

$$\left[\hat{\theta} + z_{\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}}, \hat{\theta} - z_{\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}}\right]$$

Plan

Le Jackknife

Historique

- Méthode proposée par Maurice Quenouille vers 1950 pour réduire le biais d'un estimateur.
- Vers 1958, John Tukey l'utilise pour l'estimation de l'erreur standard d'un estimateur.
- Le Jackknife préfigure le bootstrap.

Aujourd'hui le Jackknife s'utilise en complément du bootstrap, par exemple pour la construction d'intervalles de confiance.

Principe du Jackknife

Retirer une des observations de l'échantillon de départ,
puis recalculer l'estimateur sur les observations restantes.

Le Jackknife

Le sous-échantillonnage

Soit $\mathbb{X} = (X_1, \ldots, X_n)$ un n-échantillon de la loi P.
Soit $\hat{\theta} = T_n(\mathbb{X})$ un estimateur de $\theta = \theta(P)$.

Un **échantillon Jackknife** est obtenu en retirant une observation :

$$\mathbb{X}^{(-i)} = (X_1, \ldots X_{i-1}, X_{i+1}, \ldots, X_n), \quad i = 1 \ldots n$$

Une **réplique Jackknife** de $\hat{\theta}$ est l'évaluation de la statistique T_{n-1} sur un échantillon Jackknife :

$$\hat{\theta}^{(-i)} = T_{n-1}\left(\mathbb{X}^{(-i)}\right), \quad i = 1 \ldots n.$$

Le Jackknife

Réduction du biais

Soit $\mathbb{X} = (X_1, \ldots, X_n)$ un n-échantillon de P.
Soit $\hat{\theta} = T_n(\mathbb{X})$ un estimateur de $\theta = \theta(P)$.

Estimateur Jackknife du biais

$$\hat{b}_{\text{Jackk}} = (n-1)\left(\overline{\hat{\theta}^{(-)}} - \hat{\theta}\right), \quad \text{avec} \quad \overline{\hat{\theta}^{(-)}} = \frac{1}{n}\sum_{i=1}^{n} \hat{\theta}^{(-i)}$$

Estimateur Jackknife ayant un biais réduit

$$\hat{\theta}_{\text{Jackk}} = \hat{\theta} - \hat{b}_{\text{Jackk}} = n\hat{\theta} - (n-1)\overline{\hat{\theta}^{(-)}}$$

Si $\mathbb{B}(\hat{\theta}) = a/n + O(1/n^2)$, alors $\mathbb{B}(\hat{\theta}_{\text{Jackk}}) = O(1/n^2)$.

Le Jackknife

Estimation de la variance

Soit $\mathbb{X} = (X_1, \ldots, X_n)$ un n-échantillon de P.
Soit $\hat{\theta} = T_n(\mathbb{X})$ un estimateur de $\theta = \theta(P)$.

Estimateur de la variance

$$\hat{v}_{\text{Jackk}} = \frac{n-1}{n} \sum_{i=1}^{n} \left(\hat{\theta}^{(-i)} - \overline{\hat{\theta}^{(-)}} \right)^2$$

Sous certaines hypothèses,

$$\frac{\hat{v}_{\text{Jackk}}}{\mathbb{V}(\hat{\theta})} \xrightarrow{p.s.} 1.$$

Le Bootstrap "naïf" d'Efron

Le principe

Une **racine** est une fonctionnelle de l'échantillon $\mathbb{X}$ et de la loi P, notée $R_n(\mathbb{X}, P)$, dont certaines des caractéristiques (loi, ou plus simplement espérance, variance, quantile...) nous intéressent en particulier.

Principe du Bootstrap

Pour estimer la loi de $R_n(\mathbb{X}, P)$ sans faire d'hypothèse de type paramétrique sur P, Efron extrapole le principe du plug-in pour construire un monde Bootstrap miroir du monde réel dans lequel aucune quantité n'est inconnue.

Monde réel		Monde bootstrap
P	$\leftrightarrow$	$P_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$ mesure empirique
$\mathbb{X} = (X_1, \ldots, X_n)$ n-échantillon de la loi P	$\leftrightarrow$	$\mathbb{X}^* = (X_1^*, \ldots, X_n^*)$ n-échantillon de la loi P_n, appelé **échantillon bootstrap** associé à $\mathbb{X}$. $$P(X_j^* = X_i \mid \mathbb{X}) = \frac{1}{n} \ \forall i, j$$
$R_n(\mathbb{X}, P)$	$\leftrightarrow$	$R_n^* = R_n(\mathbb{X}^*, P_n)$ **réplication bootstrap** de la racine

Le Bootstrap "naïf" d'Efron

Estimation bootstrap et rééchantillonnage

Estimation bootstrap de la loi de $R_n(\mathbb{X}, P)$

La loi de $R_n(\mathbb{X}, P)$ est estimée par la loi conditionnelle de $R_n^* = R_n(\mathbb{X}^*, P_n)$ sachant $\mathbb{X}$.

Approximation de Monte Carlo et rééchantillonnage

Conditionnellement à $\mathbb{X} = (x_1, \ldots, x_n)$, une réalisation de l'échantillon bootstrap $\mathbb{X}^* = (X_1^*, \ldots, X_n^*)$ peut être simulée en tirant au hasard avec remise n valeurs dans $\{x_1, \ldots, x_n\}$.

La loi conditionnelle de R_n^* sachant $\mathbb{X}$ peut donc être approchée par une méthode de Monte Carlo.

On parle alors de rééchantillonnage.

◂ Retour à l'exemple de la moyenne

Le Bootstrap "naïf" d'Efron

Mesure et processus empiriques bootstrap

La mesure empirique et la **mesure empirique bootstrap** sont respectivement définies par

$$P_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i} \text{ et } P_n^* = \frac{1}{n}\sum_{j=1}^{n} \delta_{X_j^*}.$$

Le processus empirique et le **processus empirique bootstrap** sont respectivement définis par

$$Z_n = \sqrt{n}\,(P_n - P) \text{ et } Z_n^* = \sqrt{n}\,(P_n^* - P_n).$$

Le Bootstrap "naïf" d'Efron

Poids de rééchantillonnage

Rappel : $\mathbb{P}\left(X_j^* = X_i \middle| \mathbb{X}\right) = \frac{1}{n}$.
$\hookrightarrow$ Si $(U_1, \ldots, U_n)$ est un n-échantillon de la loi $\mathcal{U}([0,1])$, indépendant de $\mathbb{X}$, et $M_{n,i} = \sum_{j=1}^n \mathbb{1}_{U_j \in](i-1)/n, i/n]}$,

$$X_j^* =^{(\mathcal{L})} \sum_{i=1}^n X_i \mathbb{1}_{U_j \in](i-1)/n, i/n]} \text{ et } \#\{X_j^* = X_i\} =^{(\mathcal{L})} M_{n,i}.$$

Le vecteur $M_n = (M_{n,1}, \ldots, M_{n,n})$ est indépendant de $\mathbb{X}$ et de loi multinomiale de paramètres $(n, 1/n, \ldots, 1/n)$.
Les $M_{n,i}$ sont des **poids de rééchantillonnage** et M_n constitue un **plan de rééchantillonnage**.

Conditionnellement à $\mathbb{X}$,

$$P_n^* =^{(\mathcal{L})} \frac{1}{n} \sum_{i=1}^n M_{n,i} \delta_{X_i} \text{ et } Z_n^* =^{(\mathcal{L})} \frac{1}{\sqrt{n}} \sum_{i=1}^n (M_{n,i} - 1) \delta_{X_i}.$$

Le Bootstrap à poids général

Le principe

Idée : remplacer M_n par $W_n = (W_{n,1}, \ldots, W_{n,n})$ vecteur de poids (indépendants de $\mathbb{X}$) tels que $\sum W_{n,i} = n$ ou $\mathbb{E}[W_{n,i}] = 1$.

- $P_n^w = \frac{1}{n} \sum_{i=1}^n W_{n,i} \delta_{X_i}$
- $Z_n^w = c_n^w \left(P_n^w - \overline{W_n} P_n \right) = \frac{c_n^w}{n} \sum_{i=1}^n (W_{n,i} - \overline{W_n}) \delta_{X_i}$.

Racine de la forme $R_n(\mathbb{X}, P) = P_n(f)$ ou $R_n(\mathbb{X}, P) = Z_n(f)$.

Estimation bootstrap à poids de la loi de $R_n(\mathbb{X}, P)$

La loi de $R_n(\mathbb{X}, P)$ est estimée par la loi conditionnelle de $R_n^w = P_n^w(f)$ ou $Z_n^w(f)$ sachant $\mathbb{X}$.

Le Bootstrap à poids général

Exemples

Delete-d jackknife, m_n out of n bootstrap sans remise

Poids aléatoires générés par permutation de poids déterministes.

- $w = (w_1, \ldots, w_n)$ tels que $\sum_{i=1}^{n} w_i = n$.
- π_n une permutation aléatoire uniformément distribuée sur Π_n, indépendante de $\mathbb{X}$.
- $W_{n,i} = w_{\pi_n(i)}$.

Cas particulier 1 : $w = (n/(n-d), \ldots, n/(n-d), 0, \ldots, 0)$, avec un bloc de 0 de taille d, où d est supposé fixé $\hookrightarrow$ delete-d jackknife.
Cas particulier 2 : $w = (n/m_n, \ldots, n/m_n, 0, \ldots, 0)$, avec un bloc de 0 de taille $n - m_n$, où $m_n \to \infty$, $n \to \infty \hookrightarrow m_n$ out of n bootstrap sans remise.

Cas particulier du m_n out of n bootstrap sans remise.

Soit π_n une permutation aléatoire uniformément distribuée sur Π_n, et $\mathbb{X}^w = (X_1^w, \ldots, X_{m_n}^w)$, avec $X_i^w = X_{\pi_n(i)}$ pour tout $i = 1 \ldots n$.

Conditionnellement à $\mathbb{X}$, $P_n^w =^{(\mathcal{L})} \frac{1}{m_n} \sum_{i=1}^{m_n} \delta_{X_i^w}$, et

$$Z_n^w := \sqrt{\tfrac{nm_n}{n-m_n}}(P_n^w - P_n) =^{(\mathcal{L})} \sqrt{\tfrac{nm_n}{n-m_n}} \left(\tfrac{1}{m_n} \sum_{i=1}^{m_n} \delta_{X_i^w} - P_n \right).$$

Approximation de Monte Carlo et sous-échantillonnage (sans remise)

Conditionnellement à $\mathbb{X} = (x_1, \ldots, x_n)$, une réalisation de $\mathbb{X}^w = (X_1^w, \ldots, X_{m_n}^w)$ peut être simulée en tirant au hasard SANS remise m_n valeurs dans $\{x_1, \ldots, x_n\}$.

La loi conditionnelle de $R_n^w = P_n^w(f)$ ou $Z_n^w(f)$ sachant $\mathbb{X}$ peut donc être approchée par une méthode de Monte Carlo.

On parle alors de sous-échantillonnage (sans remise).

m_n out of n bootstrap avec remise

- $M_{m_n} = (M_{m_n,1}, \ldots, M_{m_n,n})$ de loi $\mathcal{M}(m_n, 1/n, \ldots, 1/n)$, indépendant de $\mathbb{X}$,
- $W_n = (W_{n,1}, \ldots, W_{n,n})$ avec $W_{n,i} = (n/m_n)M_{m_n,i}$.

Soit $\mathbb{X}^w = (X_1^w, \ldots, X_{m_n}^w)$ un m_n-échantillon de la loi P_n i.e. $P(X_j^w = X_i|\mathbb{X}) = \frac{1}{n}\ \forall i, j$.

Conditionnellement à $\mathbb{X}$, $P_n^w =^{(\mathcal{L})} \frac{1}{m_n}\sum_{i=1}^{m_n} \delta_{X_i^w}$ et

$Z_n^w := \sqrt{m_n}(P_n^w - P_n) =^{(\mathcal{L})} \sqrt{m_n}\left(\frac{1}{m_n}\sum_{i=1}^{m_n} \delta_{X_i^w} - P_n\right)$.

Approximation de Monte Carlo et sous-échantillonnage avec remise

Conditionnellement à $\mathbb{X} = (x_1, \ldots, x_n)$, une réalisation de $\mathbb{X}^w = (X_1^w, \ldots, X_{m_n}^w)$ peut être simulée en tirant au hasard AVEC remise m_n valeurs dans $\{x_1, \ldots, x_n\}$.

La loi conditionnelle de $R_n^w = P_n^w(f)$ ou $Z_n^w(f)$ sachant $\mathbb{X}$ peut donc être approchée par une méthode de Monte Carlo.

On parle alors de sous-échantillonnage avec remise.

Bootstrap bayésien

Poids de somme égale à n, générés à partir d'échantillons i.i.d.

- $Y = (Y_1, \ldots, Y_n)$ n-échantillon d'une loi d'espérance μ, d'écart-type σ, indépendant de $\mathbb{X}$,
- $W_n = (W_{n,1}, \ldots, W_{n,n})$ avec $W_{n,i} = Y_i / \overline{Y}$.

$$P_n^w = \frac{1}{n} \sum_{i=1}^{n} \frac{Y_i}{\bar{Y}} \delta_{X_i} \text{ et } Z_n^w = \frac{\sqrt{n}\mu}{\sigma} \left(P_n^w - P_n \right).$$

✔ Bootstrap plus "lisse" que les précédents.

Cas particulier : $Y = n$-échantillon de la loi exponentielle $\mathcal{E}(1)$
$\hookrightarrow$ Bootstrap bayésien.

Wild bootstrap ou bootstrap "sauvage"

$W_n = (W_{n,1}, \ldots, W_{n,n})$ n-échantillon d'une loi d'espérance 1, d'écart-type σ.

$$P_n^w = \frac{1}{n} \sum_{i=1}^{n} W_{n,i} \delta_{X_i} \text{ et } Z_n^w = \frac{\sqrt{n}}{\sigma} \left(P_n^w - \overline{W_n} P_n \right).$$

Cas particulier 1 : $W_{n,1}, \ldots, W_{n,n}$ i.i.d. de loi $\mathcal{P}(1)$
$\hookrightarrow$ poissonisation du bootstrap "naïf" d'Efron.

Cas particulier 2 : $W_{n,i} = 2B_i$, où $B_1, \ldots, B_n$ i.i.d. de loi $\mathcal{B}(1/2)$
$\hookrightarrow$ $Z_n^w = \frac{1}{\sigma\sqrt{n}} \sum_{i=1}^{n} (2B_i - 1)(\delta_{X_i} - P_n)$, où $2B_i - 1$ est une variable de Rademacher.

Le Bootstrap à poids général

Éléments de bibliographie

Ouvrages

- Barbe, P., and Bertail, P. (1995). The weighted bootstrap.
- Politis, D. N., Romano, J. P., and Wolf, M. (1999). Subsampling.
- Van der Vaart, A., and Wellner, J. A. (1996). Weak convergence and empirical processes. Pages 353–359.

Articles

- Bickel P., Götze F., and van Zwet, W. (1997). Resampling fewer than n observations : gains, losses and remedies for losses. Statistica Sinica.
- Bretagnolle, J. (1989). Lois limites du bootstrap de certaines fonctionnelles. Ann. I.H.P.
- Lo, A. Y. (1987). A large sample study of the Bayesian bootstrap. ann. Statist.

- Mammen, E. (1992) Bootstrap, wild bootstrap, and asymptotic normality. Prob. Th. Relat. Fields.
- Politis, D. N., Romano, J. P., Wolf, M. (2001). On the asymptotic theory of subsampling. Statistica Sinica.
- Praestgaard J., and Wellner, J. A. (1996). Exchangeably weighted bootstraps of the general empirical process. Ann. Prob.
- Rubin, D. (1981). The Bayesian bootstrap. Ann. Statist.
- Shao, J., and Wu, C. F. (1989). A general theory for jackknife variance estimation. Ann. Statist.
- Wu, C. F. (1990). On the asymptotic properties of the jackknife histogram. Ann. Statist.

Autres approches bootstrap

Le bootstrap paramétrique

On considère ici un modèle paramétrique.

$P = P_\tau$, $\tau \in \mathbb{R}^d$	$\leftrightarrow$	$P_{\hat{\tau}}$, où $\hat{\tau}$ est un "bon" estimateur de τ construit sur $\mathbb{X}$
$\mathbb{X} = (X_1, \ldots, X_n)$ n-échantillon de la loi P	$\leftrightarrow$	$\widehat{\mathbb{X}} = (\widehat{X}_1, \ldots, \widehat{X}_n)$ n-échantillon de la loi $P_{\hat{\tau}}$
$R_n(\mathbb{X}, P)$	$\leftrightarrow$	$\widehat{R_n} = R_n(\widehat{\mathbb{X}}, P_{\hat{\tau}})$

Estimation bootstrap paramétrique de la loi de $R_n(\mathbb{X}, P)$

La loi de $R_n(\mathbb{X}, P)$ est estimée par la loi conditionnelle de $\widehat{R_n}$ sachant $\mathbb{X}$, que l'on peut simuler par une méthode de Monte Carlo.

Autres approches bootstrap

Smooth boostrap ou bootstrap "lisse"

On considère ici un modèle non paramétrique, où P est une loi continue sur $\mathbb{R}^d$ muni de la mesure de Lebesgue.

P	$\leftrightarrow$	$\hat{P}$, où $\hat{P}$ est un "bon" estimateur à noyau de la loi P
$\mathbb{X} = (X_1, \ldots, X_n)$ n-échantillon de la loi P	$\leftrightarrow$	$\widehat{\mathbb{X}} = (\widehat{X_1}, \ldots, \widehat{X_n})$ n-échantillon de la loi $\hat{P}$
$R_n(\mathbb{X}, P)$	$\leftrightarrow$	$\widehat{R_n} = R_n(\widehat{\mathbb{X}}, \hat{P})$

Estimation de la loi de $R_n(\mathbb{X}, P)$ par bootstrap "lisse"

La loi de $R_n(\mathbb{X}, P)$ est estimée par la loi conditionnelle de $\widehat{R_n}$ sachant $\mathbb{X}$.

Plan

Cadre non paramétrique

$\mathbb{X} = (X_1, \ldots, X_n)$ est un n-échantillon d'une loi P inconnue.

$\theta(P)$ un paramètre d'intérêt de la loi P, et $\hat{\theta} = T(\mathbb{X})$ un estimateur de $\theta(P)$, par exemple $\hat{\theta} = \theta(P_n)$ estimateur plug-in (mais pas nécessairement).

On souhaite estimer sur la base d'une observation de $\mathbb{X}$:

- Le biais de $\hat{\theta} = \mathbb{E}_P[T(\mathbb{X}) - \theta(P)]$,
- La variance de $\hat{\theta} = \mathbb{V}_P[T(\mathbb{X})]$,
- L'erreur quadratique moyenne (MSE) de $\hat{\theta} = \mathbb{E}_P[(T(\mathbb{X}) - \theta(P))^2]$.

Estimation bootstrap du biais

En théorie...

Soit $\hat{\theta} = T(\mathbb{X})$ un estimateur de $\theta = \theta(P)$

Monde réel : le biais de $T(\mathbb{X})$ est donné par $\mathbb{E}[R_n(\mathbb{X}, P)]$ avec $R_n(\mathbb{X}, P) = T(\mathbb{X}) - \theta(P)$.

Monde bootstrap : $R_n^* = R_n(\mathbb{X}^*, P_n) = T(\mathbb{X}^*) - \theta(P_n)$, où $\mathbb{X}^*$ est un échantillon bootstrap ("naïf") associé à $\mathbb{X}$.

Estimateur bootstrap du biais

L'estimateur bootstrap du biais de $\hat{\theta} = T(\mathbb{X})$ est défini par $\mathbb{E}[R_n^*|\mathbb{X}] = \mathbb{E}[T(\mathbb{X}^*) - \theta(P_n)|\mathbb{X}]$.

Approximation de Monte Carlo

Conditionnellement à $\mathbb{X} = x$, la loi de R_n^* peut être simulée, et $\mathbb{E}[R_n^*|\mathbb{X} = x]$ estimée par une méthode de Monte Carlo.

Estimation bootstrap du biais

En pratique... Monte Carlo

Algorithme Estimation bootstrap du biais

Variable

B : $entier\ assez\ grand$

Début

Pour b **variant de** 1 **à** B

Générer x^{*b} réalisation d'un échantillon bootstrap

Calculer $T(x^{*b})$ réplication bootstrap de $T(x)$

FinPour

Retourner $\frac{1}{B}\sum_{b=1}^{B} T(x^{*b}) - \theta(P_n)$

Fin

Estimation bootstrap du biais

Cas particulier où $\hat{\theta} = \theta(P_n)$ (estimateur plug-in)

Algorithme Estimation bootstrap accélérée du biais

Variable

B : *entier assez grand*

Début

Pour b **variant de** 1 **à** B

Générer $x^{*b} = (x_1^{*b}, \ldots, x_n^{*b})$

Calculer $T(x^{*b})$

Calculer $p_i^{*b} = \#\left\{j,\ x_j^{*b} = x_i\right\}/n$ pour tout i

FinPour

Calculer $p_i^* = \frac{1}{B}\sum_{b=1}^{B} p_i^{*b}$ pour tout i

Retourner $\frac{1}{B}\sum_{b=1}^{B} T(x^{*b}) - \theta(P^*)$ où $P^* = \sum_{i=1}^{n} p_i^* \delta_{x_i}$

Fin

Estimation bootstrap du biais

Illustration 1 : biais de l'estimateur plug-in de la variance

Le coin du UseR

```
biaisbst=function(B,X){
  Tbst=numeric(B);n=length(X);
  for (b in 1:B){
    s=sample(1:n,n,replace=T);
    Tbst[b]=(n-1)*var(X[s])/n };
  return(mean(Tbst)-(n-1)*var(X)/n) };
biaisbst(999,X) # Estimation bootstrap du biais
```

Utilisation du package boot :

```
library(boot)
# Création d'une fonction calculant l'estimateur
# sur l'échantillon bootstrap
myvar=function(data,indice){
  n=length(data);
  return((n-1)*var(data[indice])/n) };

varboot=boot(X,myvar,R=999);
mean(varboot$t)-varboot$t0; print(varboot)
```

$\mathbb{X}$ = n-échantillon d'une loi $\mathcal{N}(0,1)$.

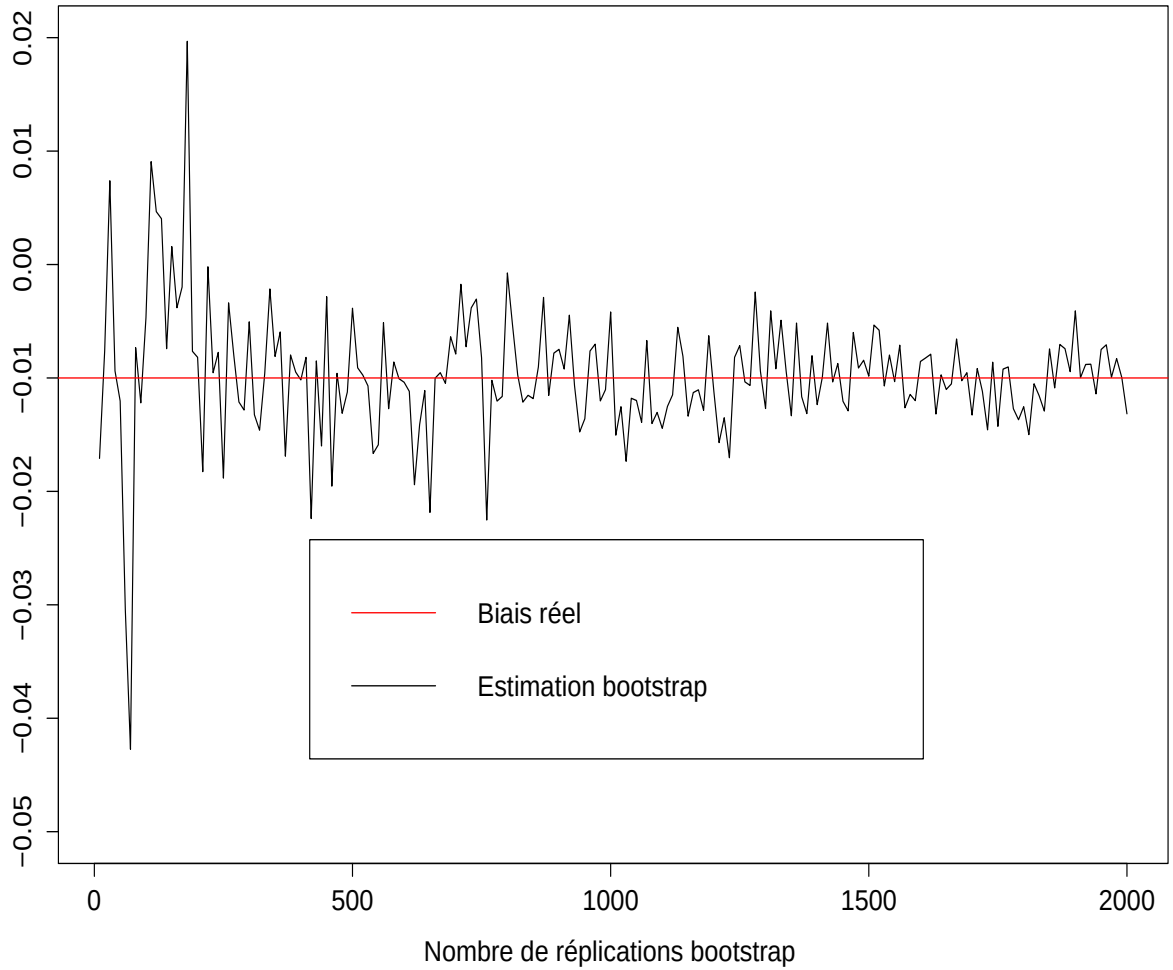

Figure: Estimation bootstrap du biais, n=100

Le coin du UseR

```
biaisbstA=function(B,X){
  Tbst=numeric(B);n=length(X);
  p=matrix(0,nr=B,nc=n);
  for (b in 1:B){
    s=sample(1:n,n,replace=T);
    Tbst[b]=(n-1)*var(X[s])/n;
    for (i in 1:n){p[b,i]=mean((X[s]==X[i]))};
    }
  pstar=apply(p,2,mean);
  return(mean(Tbst)
         - weighted.mean((X-weighted.mean(X,pstar))^2,pstar))
  };

biaisbstA(100,X) # Estimation bootstrap accélérée du biais
```

$\mathbb{X}$ = n-échantillon d'une loi $\mathcal{N}(0,1)$.

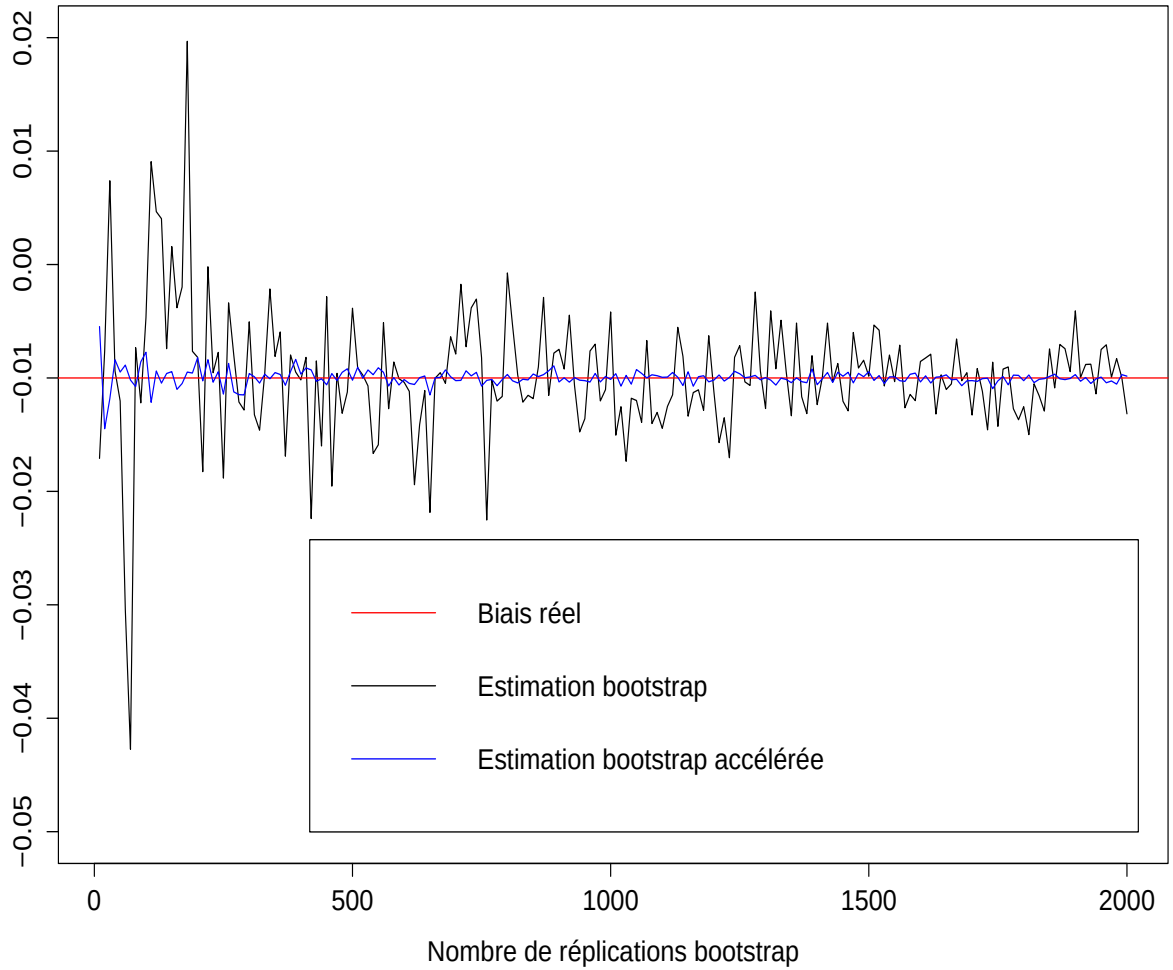

Figure: Estimation bootstrap accélérée du biais, n=100

$\mathbb{X}$ = n-échantillon d'une loi $\mathcal{N}(0,1)$.

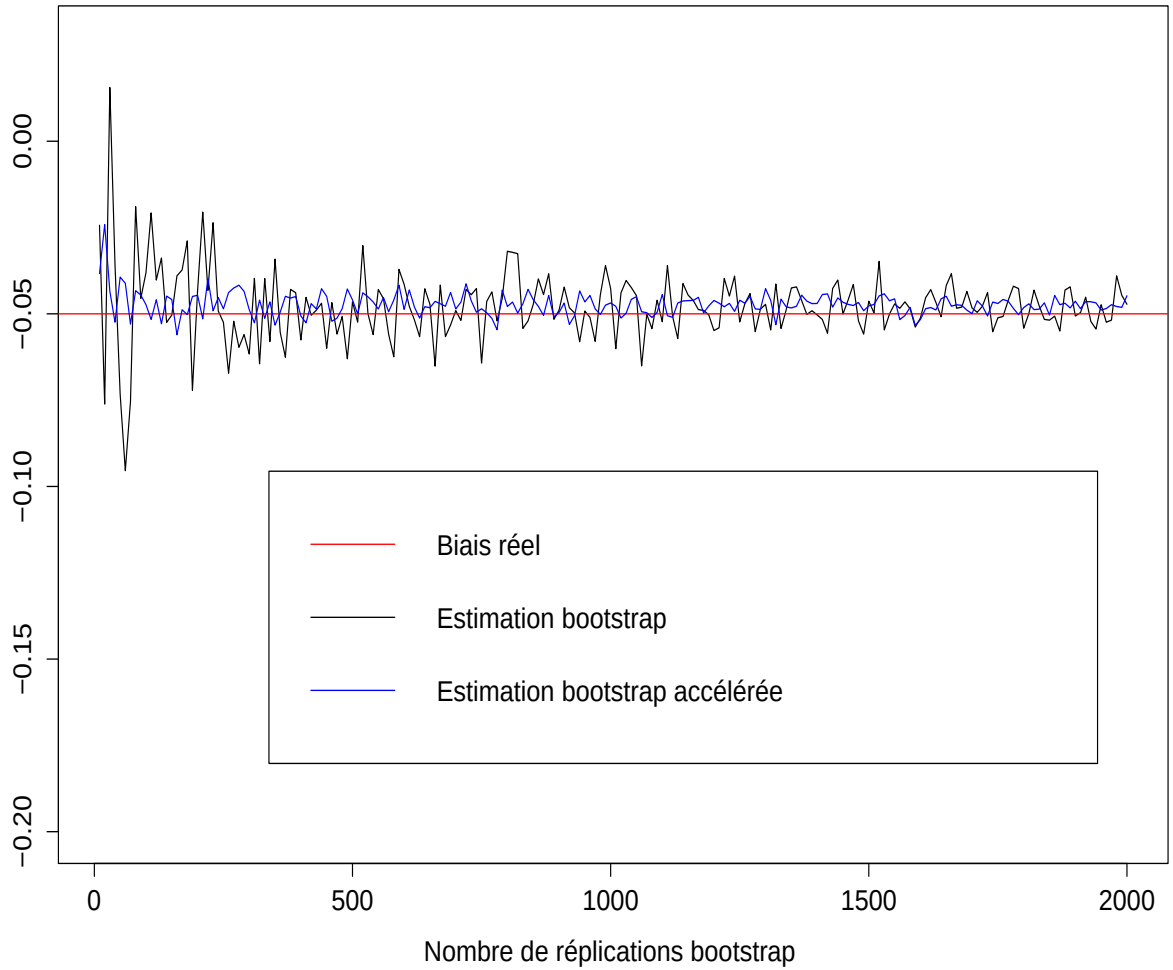

Figure: Estimation bootstrap accélérée du biais, n=20

Estimation bootstrap du biais

Illustration 2 : biais de l'estimateur plug-in du rapport des espérances

$\mathbb{X} = ((Y_1, Z_1), \ldots, (Y_n, Z_n))$ n-échantillon d'une loi bivariée P,
$\theta(P) = \mathbb{E}[Y_i]/\mathbb{E}[Z_i]$, $\hat{\theta} = \theta(P_n) = \overline{Y}/\overline{Z}$.

Le coin du UseR

```
biaisbst=function(B,X){
  Tbst=numeric(B);n=nrow(X);
  for (b in 1:B){
    s=sample(1:n,n,replace=T);
    Tbst[b]=mean(X[s,1])/mean(X[s,2]) };
  return(mean(Tbst)-mean(X[,1])/mean(X[,2]))  };
biaisbst(999,X) # Estimation bootstrap du biais

biaisbstA=function(B,X){
  Tbst=numeric(B);n=nrow(X);p=matrix(0,nr=B,nc=n);
  for (b in 1:B){
    s=sample(1:n,n,replace=T);
    Tbst[b]=mean(X[s,1])/mean(X[s,2]);
    for (i in 1:n){p[b,i]=mean(s==i)} };
  pstar=apply(p,2,mean); return(mean(Tbst)-
       weighted.mean(X[,1],pstar)/weighted.mean(X[,2],pstar)) };
biaisbstA(100,X) # Estimation bootstrap accélérée
```

$\mathbb{X} = n$-échantillon de la loi $\mathcal{E}(1) \bigotimes \mathcal{E}(2)$.

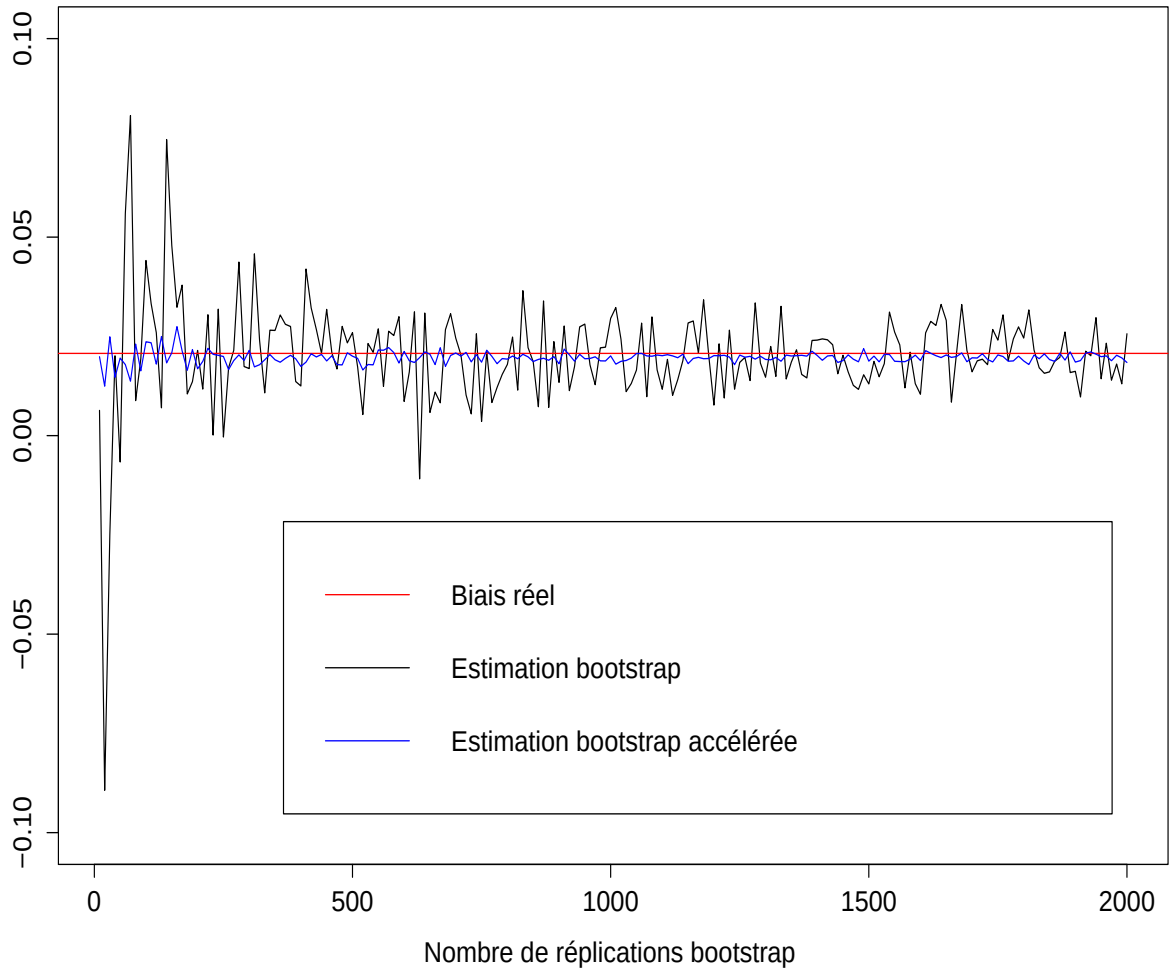

Figure: Estimation bootstrap accélérée du biais, n=100

Estimation bootstrap du biais

Choix du nombre de réplications bootstrap

Si B suffisamment grand, le TCL donne :

$$\mathbb{P}\left(\left|\frac{1}{B}\sum_{b=1}^{B} T(\mathbb{X}^{*b}) - \mathbb{E}\left[T(\mathbb{X}^{*})|\mathbb{X}\right]\right| \leq 2\sqrt{\frac{\mathbb{V}(T(\mathbb{X}^{*})|\mathbb{X})}{B}}\right) \approx 0.95.$$

L'inégalité de Markov donne :

$$\mathbb{P}\left(\left|\frac{1}{B}\sum_{b=1}^{B} T(\mathbb{X}^{*b}) - \mathbb{E}\left[T(\mathbb{X}^{*})|\mathbb{X}\right]\right| \leq \sqrt{\frac{\mathbb{V}(T(\mathbb{X}^{*})|\mathbb{X})}{0.05B}}\right) \geq 0.95.$$

✘ Problème : $\mathbb{V}(T(\mathbb{X}^{*})|\mathbb{X})$ inconnue.
✔ Préconisation : étudier graphiquement la stabilisation de l'approximation de Monte-Carlo...

Estimation bootstrap du biais

Correction du biais d'un estimateur ?

Corriger le biais de l'estimateur $\hat{\theta} = T(\mathbb{X})$ par une méthode de bootstrap revient à considérer :

$$\hat{\theta}_{corr} = \hat{\theta} - \mathbb{E}[T(\mathbb{X}^*) - \theta(P_n)|\mathbb{X}].$$

Si $\hat{\theta} = \theta(P_n)$ (estimateur plug-in),

$$\hat{\theta}_{corr} = 2\hat{\theta} - \mathbb{E}[T(\mathbb{X}^*)|\mathbb{X}].$$

✘ Attention : l'estimateur corrigé n'est pas $\mathbb{E}[T(\mathbb{X}^*)|\mathbb{X}]$ lui-même (erreur fréquente).
✘ De toutes façons, corriger le biais en pratique est dangereux, car $\hat{\theta}_{corr}$ peut avoir une plus grande variance que $\hat{\theta}$!
✔ Mais avoir une idée du biais permet de mieux arbitrer le compromis biais-variance...

Estimation bootstrap de la variance

En théorie...

Monde réel : la variance de $\hat{\theta} = T(\mathbb{X})$ est donnée par $\mathbb{V}_P[R_n(\mathbb{X}, P)]$ avec $R_n(\mathbb{X}, P) = T(\mathbb{X})$.

Monde bootstrap : $R_n^* = R_n(\mathbb{X}^*, P_n) = T(\mathbb{X}^*)$, où $\mathbb{X}^*$ est un échantillon bootstrap ("naïf") associé à $\mathbb{X}$.

Estimateur bootstrap de la variance

L'estimateur bootstrap de la variance de $\hat{\theta} = T(\mathbb{X})$ est défini par $\mathbb{V}[R_n^*|\mathbb{X}] = \mathbb{V}[T(\mathbb{X}^*)|\mathbb{X}]$.

Approximation de Monte Carlo

Conditionnellement à $\mathbb{X} = x$, la loi de R_n^* peut être simulée, et $\mathbb{V}[R_n^*|\mathbb{X} = x]$ estimée par une méthode de Monte Carlo.

Estimation bootstrap de la variance

En pratique... Monte Carlo

Algorithme Estimation bootstrap de la variance

Variable

B : *entier assez grand*

Début

Pour b **variant de** 1 **à** B

Générer x^{*b} réalisation d'un échantillon bootstrap

Calculer $T(x^{*b})$ réplication bootstrap de $T(x)$

FinPour

Retourner $\frac{1}{B-1}\sum_{b=1}^{B}\left(T(x^{*b}) - \frac{1}{B}\sum_{b=1}^{B}T(x^{*b})\right)^2$

Fin

Estimation bootstrap de la MSE

En théorie...

Monde réel : l'erreur quadratique moyenne de $\hat{\theta} = T(\mathbb{X})$ est égale à $\mathbb{E}_P[R_n^2(\mathbb{X}, P)]$ avec $R_n(\mathbb{X}, P) = T(\mathbb{X}) - \theta(P)$.

Monde bootstrap : $R_n^* = R_n(\mathbb{X}^*, P_n) = T(\mathbb{X}^*) - \theta(P_n)$, où $\mathbb{X}^*$ est un échantillon bootstrap ("naïf") associé à $\mathbb{X}$.

Estimateur bootstrap de la MSE

L'estimateur bootstrap de l'erreur quadratique moyenne de $\hat{\theta} = T(\mathbb{X})$ est défini par $\mathbb{E}[(R_n^*)^2|\mathbb{X}] = \mathbb{E}\left[(T(\mathbb{X}^*) - \theta(P_n))^2 \,\middle|\, \mathbb{X}\right]$.

Approximation de Monte Carlo

Conditionnellement à $\mathbb{X} = (x_1, \ldots, x_n)$, la loi de R_n^* peut être simulée, et $\mathbb{E}[(R_n^*)^2|\mathbb{X}]$ estimée par une méthode de Monte Carlo.

Estimation bootstrap de la MSE

En pratique... Monte Carlo

Algorithme Estimation bootstrap de la MSE

Variable

B : $entier\ assez\ grand$

Début

Pour b **variant de** 1 **à** B

Générer x^{*b} réalisation d'un échantillon bootstrap

Calculer $T(x^{*b})$ réplication bootstrap de $T(x)$

FinPour

Retourner $\frac{1}{B}\sum_{b=1}^{B}\left(T(x^{*b}) - \theta(P_n)\right)^2$

Fin

Pour résumer

$$
\left.
\begin{array}{ccccc}
 & & x^{*1} & \longrightarrow & T(x^{*1}) \\
 & \nearrow & & & \\
x & \longrightarrow & \vdots & \longrightarrow & \vdots \\
 & \searrow & & & \\
 & & x^{*B} & \longrightarrow & T(x^{*B})
\end{array}
\right\}
\begin{array}{l}
\frac{1}{B}\sum_{b=1}^{B} T(x^{*b}) - \theta(P_n) \\
\frac{1}{B-1}\sum_{b=1}^{B}\left(T(x^{*b}) - \frac{1}{B}\sum_{b=1}^{B} T(x^{*b})\right)^2 \\
\frac{1}{B}\sum_{b=1}^{B}\left(T(x^{*b}) - \theta(P_n)\right)^2
\end{array}
$$

Plan

Intervalle de confiance et pivot

Soit $\mathbb{X} = (X_1, \dots X_n)$ un n-échantillon de P.
Soit $\theta = \theta(P) \in \Theta$ un paramètre réel.
Soit $\alpha \in]0, 1[$.

Un **intervalle de confiance pour θ de niveau $1 - \alpha$**, $\mathrm{IC}_{1-\alpha}(\theta)$, est un sous-ensemble aléatoire,

$$\mathcal{I}_n(\mathbb{X}) = [T_{\inf}(\mathbb{X}), T_{\sup}(\mathbb{X})],$$

de Θ tel que

$$1 - \alpha = \mathbb{P}\left(\theta \in \mathcal{I}_n(\mathbb{X})\right).$$

Construction d'un $\mathrm{IC}_{1-\alpha}(\theta)$ par la méthode du pivot

- Trouver un pivot $R_n(\mathbb{X}, \theta)$,
- Déterminer des quantiles $q_{\alpha/2}$ et $q_{1-\alpha/2}$ tels que

$$\mathbb{P}\Big(R_n(\mathbb{X}, \theta) \in [q_{\alpha/2}, q_{1-\alpha/2}]\Big) = 1 - \alpha,$$

- Inverser l'équation :

$$R_n(\mathbb{X}, \theta) \in [q_{\alpha/2}, q_{1-\alpha/2}] \quad \Leftrightarrow \quad \theta \in \Big[T_{\inf}(\mathbb{X}), T_{\sup}(\mathbb{X})\Big]$$

Construction d'un $\mathrm{IC}_{1-\alpha}(\theta)$ par pivotement de la f.d.r

Pour tout $\alpha \in]0,1[$, il existe une statistique $\hat{\theta}_\alpha$ telle que :

$$\alpha = \mathbb{P}\left(\theta < \hat{\theta}_\alpha\right).$$

Alors $\left[\hat{\theta}_{\alpha/2}(\mathbb{X}), \hat{\theta}_{1-\alpha/2}(\mathbb{X})\right]$ est un $\mathrm{IC}_{1-\alpha}(\theta)$.

Les statistiques $\hat{\theta}_\alpha$ peuvent être obtenues en pivotant la f.d.r. d'une statistique T.

Intervalle de confiance par tables bootstrap ou bootstrap-t

Soit $R_n(\mathbb{X}, \theta)$ un pivot (asymptotique)

Monde réel : $q_{\alpha/2}$ et $q_{1-\alpha/2}$ les quantiles d'ordre $\alpha/2$ et $1-\alpha/2$ de la loi de $R_n(\mathbb{X}, \theta)$,

$$\mathrm{IC}_{1-\alpha}(\theta) = \left\{ t \in \Theta : \ R_n(\mathbb{X}, t) \in [q_{\alpha/2}, q_{1-\alpha/2}] \right\}$$

Monde bootstrap : $q^*_{\alpha/2}$ et $q^*_{1-\alpha/2}$ les quantiles d'ordre $\alpha/2$ et $1-\alpha/2$ de la loi conditionnelle de $R^*_n = R_n(\mathbb{X}^*, \theta(P_n))$ sachant $\mathbb{X}$

$$\mathrm{IC}^*_{1-\alpha}(\theta) = \left\{ t \in \Theta : \ R_n(\mathbb{X}, t) \in [q^*_{\alpha/2}, q^*_{1-\alpha/2}] \right\}$$

✘ Existence d'un pivot, en pratique estimateurs supposés asymptotiquement normaux

✘ Non invariante par transformation

Algorithme IC bootstrap-t

$\{$ *Évaluer* $q^*_{\alpha/2}$ *et* $q^*_{1-\alpha/2}$ $\}$

Variable

B : 999, 9999...

Début

Pour b **variant de** 1 **à** B

Générer x^{*b} réalisation d'un échantillon bootstrap

Calculer $r_n^{*b} = R_n(x^{*b}, \theta(P_n))$ réplication bootstrap de R_n

FinPour

Retourner les stat. d'ordre $\lceil B\alpha/2 \rceil$ et $\lceil B(1-\alpha/2) \rceil$ de $(r_n^{*b})_b$

Fin

Exemple : θ est un paramètre de position pour $\hat{\theta}$

- $R = \hat{\theta} - \theta$ pivot
- $\hat{\theta} - \theta \sim \mathcal{N}(0, \sigma_n^2)$ approximation normale
- $R = \left(\hat{\theta} - \theta\right) / \hat{\sigma}_n$ pivot
- $R = \beta(\hat{\theta}) - \beta(\theta)$ stabilisation de la variance

Illustration : estimation par intervalle de confiance de la variance correspondant aux mesures de la longueur en pouces de l'avant-bras de 140 hommes adultes.

Estimateur :

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$$

Racine :

$$R_n = \left(\hat{\sigma}^2 - \sigma^2\right)/\mathbb{V}(\hat{\sigma}^2).$$

Résultat :

$$[1.01 \quad 1.47]$$

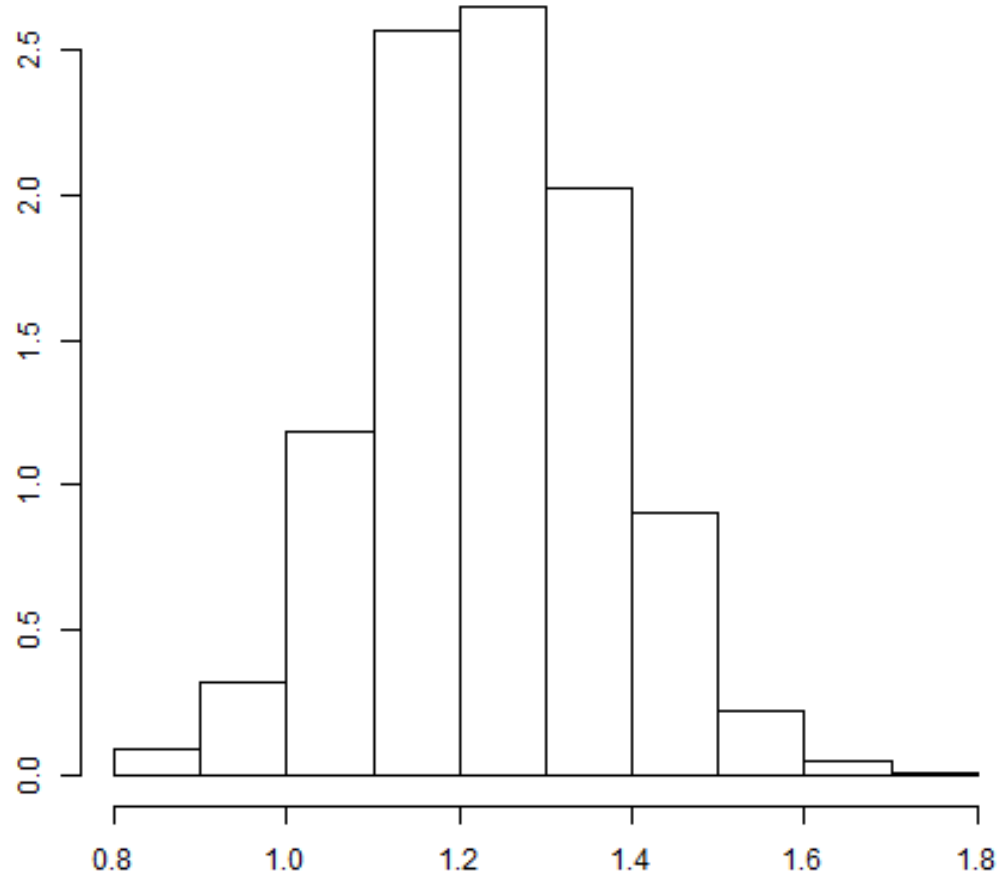

Figure: Histogramme des réplications bootstrap de $\hat{\sigma}^2$

Intervalle de confiance par percentile bootstrap

Soit $\hat{\theta} = \theta(P_n) = T(\mathbb{X})$ un estimateur plug-in de θ.
Soit $\beta : \Theta \to \mathbb{R}$ bijective croissante telle que :

$$\mathbb{P}\left(\beta(\hat{\theta}) - \beta(\theta) \leq x\right) = \psi(x),$$

où ψ la f.d.r d'une loi continue symétrique.

Monde réel : q_α le quantile d'ordre α de ψ,

$$\mathrm{IC}_{1-\alpha}(\theta) = \left[\beta^{-1}\left(\beta(\hat{\theta}) + q_{\alpha/2}\right), \beta^{-1}\left(\beta(\hat{\theta}) + q_{1-\alpha/2}\right)\right]$$

Monde bootstrap t^*_α le quantile d'ordre α de la loi de $T(\mathbb{X}^*)|\mathbb{X}$.

$$\mathrm{IC}^*_{1-\alpha}(\theta) = \left[t^*_{\alpha/2}, t^*_{1-\alpha/2}\right]$$

✔ Facile à implémenter, invariante par transformation

✘ Niveau de confiance approximatif

Algorithme IC percentile bootstrap

$\{$ *Évaluer* $t^*_{\alpha/2}$ *et* $t^*_{1-\alpha/2}$ $\}$

Variable

B : 999, 9999...

Début

Pour b **variant de** 1 **à** B

Générer x^{*b} réalisation d'un échantillon bootstrap

Calculer $t^{*b} = T(x^{*b})$ réplication bootstrap de $\hat{\theta}$

FinPour

Retourner les stat. d'ordre $\lceil B\alpha/2 \rceil$ et $\lceil B(1-\alpha/2) \rceil$ de $(t^{*b})_b$

Fin

Intervalle de confiance BC-percentile bootstrap

Extension de la méthode percentile bootstrap au cas d'un estimateur biaisé :

$$\mathbb{P}\left(\beta(\hat{\theta}) - \beta(\theta) + b_0 \leq x\right) = \psi(x),$$

où b_0 est le biais de $\hat{\theta}$.

Monde réel : q_α le quantile d'ordre α de ψ,

$$\mathrm{IC}_{1-\alpha}(\theta) = \left[\beta^{-1}\left(\beta(\hat{\theta}) + b_0 + q_{\alpha/2}\right), \beta^{-1}\left(\beta(\hat{\theta}) + b_0 + q_{1-\alpha/2}\right)\right]$$

Monde bootstrap : t^*_α le quantile d'ordre α de la loi de $T(\mathbb{X}^*)|\mathbb{X}$,
$\alpha_1 = \psi(2b^*_0 + q_{\alpha/2})$, $\alpha_2 = \psi(2b^*_0 + q_{1-\alpha/2})$,
$b^*_0 = \psi^{-1}\left(\mathbb{P}\left(T(\mathbb{X}^*) < \hat{\theta}|\mathbb{X}\right)\right)$.

$$\mathrm{IC}^*_{1-\alpha}(\theta) = \left[t^*_{\alpha_1}, t^*_{\alpha_2}\right]$$

Remarque : on choisit pour ψ la f.d.r de $\mathcal{N}(0,1)$

Algorithme IC BC-percentile bootstrap

$\{$ *Évaluer* $t^*_{\alpha_1}$ *et* $t^*_{\alpha_2}$ $\}$

Variable

B : 999, 9999...

Début

Pour b **variant de** 1 **à** B

Générer x^{*b} réalisation d'un échantillon bootstrap

Calculer $t^{*b} = T(x^{*b})$ réplication bootstrap de $\hat{\theta}$

FinPour

$\hat{b}^*_0 = \psi^{-1}\left(\frac{1}{B}\sum_b \mathbb{1}_{t^{*b}\leq\hat{\theta}}\right)$

$\hat{\alpha}^*_1 = \psi\left(2\hat{b}^*_0 + q_{\alpha/2}\right)$, $\hat{\alpha}^*_2 = \psi\left(2\hat{b}^*_0 + q_{1-\alpha/2}\right)$

Retourner les stat. d'ordre $\lceil B\hat{\alpha}^*_1 \rceil$ et $\lceil B\hat{\alpha}^*_2 \rceil$ de $(t^{*b})_b$

Fin

Intervalle de confiance BC_a-percentile bootstrap

Extension de la méthode BC-percentile au cas d'un estimateur avec une curtose.

$$\mathbb{P}\left(\frac{\beta(\hat{\theta})-\beta(\theta)}{1+a\beta(\theta)}+b_0\leq x\right)=\psi(x),$$

où a est une constante d'accélération

Monde réel : $q_{\alpha/2}$ le quantile d'ordre $\alpha/2$ de ψ,

$$\mathrm{IC}_{1-\alpha}(\theta)=\left[\beta^{-1}\left(\beta(\hat{\theta})+\frac{(1+a\beta(\hat{\theta}))(b_0+q_{\alpha/2})}{1-a(b_0+q_{\alpha/2})}\right),\beta^{-1}\left(\beta(\hat{\theta})+\frac{(1+a\beta(\hat{\theta}))(b_0+q_{1-\alpha/2})}{1-a(b_0+q_{1-\alpha/2})}\right)\right]$$

Monde bootstrap : $t^*_{\alpha/2}$ le quantile d'ordre $\alpha/2$ de la loi de $T(\mathbb{X}^*)|\mathbb{X}$, $\alpha_1=\psi\left(b_0^*+\frac{b_0^*+q_{\alpha/2}}{1-a(b_0^*+q_{\alpha/2})}\right)$, $\alpha_2=\psi\left(b_0^*+\frac{b_0^*+q_{1-\alpha/2}}{1-a(b_0^*+q_{1-\alpha/2})}\right)$.

$$\mathrm{IC}^*_{1-\alpha}(\theta)=\left[t^*_{\alpha_1},t^*_{\alpha_2}\right]$$

✘ Estimateur bootstrap de a : dépend du modèle

Plan

Tests de permutation

Principe général

Fisher (1935) / Hoeffding (1952), Romano et Wolf (2005)

Soit $\mathbb{X}$ une v.a. modélisant les données observées, à valeurs dans $\mathcal{X}$.

(H_0) La loi de $\mathbb{X}$ appartient à une famille de lois $\mathcal{P}$

Statistique de test : $T(\mathbb{X})$ à valeurs réelles, dont la loi sous (H_0) est inconnue ou n'est pas libre, pour laquelle de grandes valeurs conduisent au rejet de (H_0).

Soit $\mathcal{G}$ un groupe fini de transformations : $\mathcal{X} \to \mathcal{X}$ de cardinal M.

Hypothèse d'invariance

Sous (H_0), la loi de $\mathbb{X}$ est invariante par les transformations $g \in \mathcal{G}$, i.e. pour tout $g \in \mathcal{G}$, $g\mathbb{X} =^{(\mathcal{L})} \mathbb{X}$.

Sachant $\mathbb{X} = x$, soit $T^{(1)}(x) \leq \ldots \leq T^{(M)}(x)$ les valeurs ordonnées de $T(gx),\ g \in \mathcal{G}$.

$a(x) = \frac{M\alpha - M^+(x)}{M^0(x)}$,

$M^+(x) = \#\left\{j,\ T^{(j)}(x) > T^{(\lceil M - M\alpha \rceil)}(x)\right\}$,

$M^0(x) = \#\left\{j,\ T^{(j)}(x) = T^{(\lceil M - M\alpha \rceil)}(x)\right\}$.

Test randomisé

Soit $\alpha \in]0,1[$. Le test randomisé défini par :

$$\Phi_\alpha(\mathbb{X}) = \begin{cases} 1 & \text{si} \quad T(\mathbb{X}) > T^{(\lceil M - M\alpha \rceil)}(\mathbb{X}) \\ a(\mathbb{X}) & \text{si} \quad T(\mathbb{X}) = T^{(\lceil M - M\alpha \rceil)}(\mathbb{X}) \\ 0 & \text{si} \quad T(\mathbb{X}) < T^{(\lceil M - M\alpha \rceil)}(\mathbb{X}) \end{cases}$$

est de niveau α, et la p-valeur de Φ_α en x est donnée par :

$$p(x) = \frac{1}{M} \#\left\{g \in \mathcal{G},\ T(gx) \geq T(x)\right\}.$$

Tests non-randomisés approchés

Soit x une observation de $\mathbb{X}$.

- Soit $g_1, \ldots, g_B$ i.i.d. de loi uniforme sur $\mathcal{G}$ et $t^{*b} = T(g_b x)$. Soit $(t^{*(1)}, \ldots, t^{*(B)})$ la statistique d'ordre associée à $(t^{*1}, \ldots, t^{*B})$.
  $T^{(\lceil M - M\alpha \rceil)}(x)$ est approchée par $t^{*(\lceil B - B\alpha \rceil)}$ (Monte Carlo).
- Premier test : on rejette (H_0) si $T(x) > t^{*(\lceil B - B\alpha \rceil)}$ (niveau contrôlé par $\frac{\lfloor B\alpha \rfloor + 1}{B+1}$).
- Deuxième test : on rejette (H_0) si la p-valeur approchée $p^*_B(x) = \frac{\#\{b, t^{*b} \geq T(x)\} + 1}{B+1}$ est inférieure ou égale à α (niveau contrôlé par α).

Tests de permutation

Problème à deux échantillons ou comparaison de lois

$\mathbb{X}^{(1)} = (X_1^{(1)}, \ldots, X_{n_1}^{(1)})$ un n_1-échantillon d'une loi P_1 inconnue,
$\mathbb{X}^{(2)} = (X_1^{(2)}, \ldots, X_{n_2}^{(2)})$ un n_2-échantillon d'une loi P_2 inconnue,
supposés indépendants.

(H_0) $P_1 = P_2$ versus (H_1) $P_1 \neq P_2$

Échantillon agrégé : $n = n_1 + n_2$,
$\mathbb{X} = (X_1, \ldots, X_n) := (X_1^{(1)}, \ldots, X_{n_1}^{(1)}, X_1^{(2)}, \ldots, X_{n_2}^{(2)})$.

Statistique de test : $T(\mathbb{X})$ à valeurs réelles, dont la loi sous (H_0) est inconnue ou n'est pas libre de $P_1 = P_2$, pour laquelle de grandes valeurs conduisent au rejet de (H_0) (par exemple, la statistique de Kolmogorov-Smirnov généralisée).

Groupe d'invariance :
$\mathcal{G} = \{g : x = (x_1, \ldots, x_n) \mapsto (x_{\pi(1)}, \ldots, x_{\pi(n)}),\ \pi \in \Pi_n\}$,
Π_n =l'ensemble des permutations de $\{1, \ldots, n\}$. $M = n!$

Algorithme Test de permutation d'égalité des lois

Variable

B : 999, 9999... *tel que* $(B+1)\alpha$ *entier*

Début

Pour b **variant de** 1 **à** B

Générer $\pi_b \sim \mathcal{U}(\Pi_n)$

Calculer $g_b x = (x_{\pi_b(1)}, \ldots, x_{\pi_b(n)})$

Calculer $t^{*b} = T(g_b x)$

FinPour

Retourner $p_B^*(x) = \frac{\#\{b, t^{*b} \geq T(x)\}+1}{B+1}$

Fin

Exemple

X_i à valeurs réelles, $T(\mathbb{X}) = \left| \frac{1}{n_1} \sum_{i=1}^{n_1} X_i - \frac{1}{n_2} \sum_{i=n_1+1}^{n_2} X_i \right|$.

Le coin du UseR

```
myT=function(data,i,n1,n2){
  xstar=data[i];
  return(abs(mean(xstar[1:n1])-mean(xstar[(n1+1):(n1+n2)])))
  };
pval=function(X1,X2,B){
  n1=length(X1);n2=length(X2);
  X=c(X1,X2);
  Tbst=boot(X,myT,R=B,sim="permutation",n1=n1,n2=n2);
  return(c(Tbst$t0,(1+sum(Tbst$t>=Tbst$t0))/(B+1)))}
  };
```

Illustration

$P_1 = \mathcal{N}(1,4)$, $n_1 = 30$, $P_2 = \mathcal{N}(0,1)$, $n_2 = 35$, $T(x) = 0.988$,
$p^*_{9999}(x) = 0.03397$.

Tests de permutation

Test d'indépendance

(Y, Z) couple de variables aléatoires de loi P
$\mathbb{X} = ((Y_1, Z_1), \ldots, (Y_n, Z_n))$ $n-$échantillon de la loi P.

(H_0) "Y et Z sont indépendantes" versus
(H_1) "Y et Z ne sont indépendantes"

Statistique de test : $T(\mathbb{X})$ à valeurs réelles, dont la loi sous (H_0) est inconnue, pour laquelle de grandes valeurs conduisent au rejet de (H_0).

Groupe d'invariance : $\mathcal{G} = \{g : x = ((y_1, z_1), \ldots, (y_n, z_n)) \mapsto ((y_1, z_{\pi(1)}), \ldots, (y_n, z_{\pi(n)})),\ \pi \in \Pi_n\}$. $M = n!$

Algorithme Test de permutation d'indépendance

Variable

B : 999, 9999... *tel que* $(B+1)\alpha$ *entier*

Début

Pour b **variant de** 1 **à** B

Générer $\pi_b \sim \mathcal{U}(\Pi_n)$

Calculer $g_b x = ((y_1, z_{\pi_b(1)}), \ldots, (y_n, z_{\pi_b(n)}))$

Calculer $t^{*b} = T(g_b x)$

FinPour

Retourner $p^*_B(x) = \frac{\#\{b, t^{*b} \geq T(x)\}+1}{B+1}$

Fin

Exemple

$T(\mathbb{X})$ = coefficient de corrélation de Pearson.

Le coin du UseR

```
myT=function(data,i,n){
  xstar=data[i];
  return(cor(xstar[1:n],xstar[(n+1):(2*n)]))  };
pval=function(XX,B){
  n=nrow(XX);X=c(XX[,1],XX[,2]);
  Tbst=boot(X,myT,R=B,sim="permutation",n=n,strata=rep(c(1,2),c(n,n)));
  return(list(Tbst$t0,(1+sum(Tbst$t>=Tbst$t0))/(B+1))) };
```

Illustration

Données : espérance de vie, QI moyen, consommation d'alcool moyenne par habitant pour tous les pays.

- Espérance de vie et QI moyen :
  $T(x) = 0.8488$, $p^*_{9999}(x) = 1.10^{-4}$.
- Espérance de vie et consommation moyenne d'alcool :
  $T(x) = 0.3058$, $p^*_{9999}(x) = 6.10^{-4}$.

Tests bootstrap

Problème à deux échantillons

$\mathbb{X}^{(1)} = (X_1^{(1)}, \ldots, X_{n_1}^{(1)})$ un n_1-échantillon d'une loi P_1 inconnue,
$\mathbb{X}^{(2)} = (X_1^{(2)}, \ldots, X_{n_2}^{(2)})$ un n_2-échantillon d'une loi P_2 inconnue,
supposés indépendants.

(H_0) $P_1 = P_2$ versus (H_1) $P_1 \neq P_2$.

Échantillon agrégé : $n = n_1 + n_2$,
$\mathbb{X} = (X_1, \ldots, X_n) := (X_1^{(1)}, \ldots, X_{n_1}^{(1)}, X_1^{(2)}, \ldots, X_{n_2}^{(2)})$.

Statistique de test : $T(\mathbb{X})$ à valeurs réelles, dont la loi sous (H_0) est inconnue ou n'est pas libre de $P_1 = P_2$, pour laquelle de grandes valeurs conduisent au rejet de (H_0).

Soit $\mathbb{X}^*$ un échantillon bootstrap associé à $\mathbb{X}$.

Test bootstrap

Soit $\alpha \in]0,1[$. Le test bootstrap associé à $T(\mathbb{X})$ est donné par $\Phi_\alpha * (\mathbb{X}) = 1$ si $T(\mathbb{X}) > q^*_{1-\alpha}(\mathbb{X})$, 0 sinon, où $q^*_{1-\alpha}(\mathbb{X})$ est le $(1-\alpha)$ quantile de la loi de $T(\mathbb{X}^*)$ sachant $\mathbb{X}$.

Tests bootstrap approchés

- Soit $\mathbb{X}^{*1}, \ldots, \mathbb{X}^{*B}$ B échantillons bootstrap.
- Soit $T^{*b} = T(\mathbb{X}^{*b})$.
- Soit $(T^{*(1)}, \ldots, T^{*(B)})$ statist. d'ordre associée à $(T^{*1}, \ldots, T^{*B})$.

Le quantile conditionnel $q^*_{1-\alpha}(\mathbb{X})$ est approché par $T^{*(\lceil B - B\alpha \rceil)}$ (Monte Carlo).

- Premier test : on rejette (H_0) si $T(\mathbb{X}) > T^{*(\lceil B - B\alpha \rceil)}$.
- Deuxième test : étant donnée une observation x de $\mathbb{X}$, on rejette (H_0) si la p-valeur approchée $p^*_B(x) = \frac{\#\{b,\ t^{*b} \geq T(x)\}+1}{B+1}$ est inférieure ou égale à α.

Algorithme Test bootstrap d'égalité des lois

Variable

B : 999, 9999... *tel que* $(B+1)\alpha$ *entier*

Début

Pour b **variant de** 1 **à** B

Générer x^{*b} réalisation d'un échantillon bootstrap

Calculer $t^{*b} = T(x^{*b})$ réplique bootstrap de $T(x)$

FinPour

Retourner $p^*_B(x) = \frac{\#\{b,\ t^{*b} \geq T(x)\}+1}{B+1}$

Fin

Exemple

X_i sont à valeurs réelles, $T(\mathbb{X}) = \left| \frac{1}{n_1} \sum_{i=1}^{n_1} X_i - \frac{1}{n_2} \sum_{i=n_1+1}^{n_2} X_i \right|$.

Le coin du UseR

```
myT=function(data,i,n1,n2){
  xstar=data[i];
  return(abs(mean(xstar[1:n1])-mean(xstar[(n1+1):(n1+n2)])))  };
pval=function(X1,X2,B){
  n1=length(X1);n2=length(X2);X=c(X1,X2);
  Tbst=boot(X,myT,R=B,sim="ordinary",n1=n1,n2=n2);
  return(list(Tbst$t0,(1+sum(Tbst$t>=Tbst$t0))/(B+1))) };
```

Illustration (Même exemple que pour le test de permutation)

$P_1 = \mathcal{N}(1,4)$, $n_1 = 30$, $P_2 = \mathcal{N}(0,1)$, $n_2 = 35$, $T(x) = 0.988$,
$p^*_{9999}(x) = 0.014985$ (versus 0.03397 pour le test de permutation).

Tests bootstrap

Test d'indépendance

(Y, Z) un couple de variables aléatoires de loi P
$\mathbb{X} = ((Y_1, Z_1), \ldots, (Y_n, Z_n))$ un $n-$échantillon de la loi P.
$\mathbb{Y} = (Y_1, \ldots, Y_n)$ et $\mathbb{Z} = (Z_1, \ldots, Z_n)$.

(H_0) "Y et Z sont indépendantes" versus
(H_1) "Y et Z ne sont indépendantes" .

Statistique de test : $T(\mathbb{Y}, \mathbb{Z})$ à valeurs réelles, dont la loi sous (H_0) est inconnue, pour laquelle de grandes valeurs conduisent au rejet de (H_0).

$\mathbb{Y}^*$ échantillon bootstrap associé à $\mathbb{Y}$,
$\mathbb{Z}^*$ échantillon bootstrap associé à $\mathbb{Z}$.

Test bootstrap

Soit $\alpha \in]0,1[$. Le test bootstrap associé à $T(\mathbb{Y},\mathbb{Z})$ est donné par $\Phi_\alpha*(\mathbb{Y},\mathbb{Z}) = 1$ si $T(\mathbb{Y},\mathbb{Z}) > q^*_{1-\alpha}(\mathbb{Y},\mathbb{Z})$, 0 sinon, où $q^*_{1-\alpha}(\mathbb{Y},\mathbb{Z})$ est le $(1-\alpha)$ quantile de la loi de $T(\mathbb{Y}^*,\mathbb{Z}^*)$ sachant $\mathbb{X}$.

Algorithme Test bootstrap dindépendance

Variable

B : 999, 9999... *tel que* $(B+1)\alpha$ *entier*

Début

Pour b **variant de** 1 **à** B

Générer y^{*b} et z^{*b} échantillons bootstrap associés à y et z

Calculer $t^{*b} = T(y^{*b}, z^{*b})$

FinPour

Retourner $p^*_B(y,z) = \frac{\#\{b,\ t^{*b} \geq T(y,z)\}+1}{B+1}$

Fin

Exemple
Le coin du UseR

```
myT=function(data,i,n){
  xstar=data[i];
  return(cor(xstar[1:n],xstar[(n+1):(2*n)])) };
pval=function(XX,B){
  n=nrow(XX);X=c(XX[,1],XX[,2]);
  Tbst=boot(X,myT,R=B,sim="ordinary",n=n,strata=rep(c(1,2),c(n,n)));
  return(list(Tbst$t0,(1+sum(Tbst$t>=Tbst$t0))/(B+1))) };
```

Illustration

Données : espérance de vie, QI moyen, consommation d'alcool moyenne par habitant pour tous les pays.

- Espérance de vie et QI moyen :
  $T(x) = 0.8488$, $p^*_{9999}(x) = 1.10^{-4}$.
- Espérance de vie et consommation moyenne d'alcool :
  $T(x) = 0.3058$, $p^*_{9999}(x) = 4.10^{-4}$.

Tests bootstrap

Test sur la moyenne

$\mathbb{X} = (X_1, \ldots, X_n)$ n-échantillon d'une loi P sur $\mathbb{R}$, d'espérance $\theta(P) = \int x dP(x)$, telle que $\int x^2 dP(x) < \infty$.

(H_0) $\theta(P) = \theta_0$ versus (H_1) $\theta(P) > \theta_0$.

Statistique de test "studentisée" :
$T(\mathbb{X}) = \frac{\sqrt{n}\left(\overline{\mathbb{X}} - \theta_0\right)}{S(\mathbb{X})}$, où $S^2(\mathbb{X}) = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{\mathbb{X}})^2$.

Test asymptotique

Soit $\alpha \in]0,1[$.
Le test défini par $\Phi_\alpha(\mathbb{X}) = \mathbb{1}_{\{T(\mathbb{X}) > q_{1-\alpha}\}}$, où $q_{1-\alpha}$ est le $(1-\alpha)$ quantile de la loi $\mathcal{N}(0,1)$, est de niveau asymptotique α.

Soit $R_n(\mathbb{X}, P) = \sqrt{n}\frac{\overline{\mathbb{X}}-\theta(P)}{S(\mathbb{X})}$ et $R_n^* = R_n(\mathbb{X}^*, P_n) = \sqrt{n}\frac{\overline{\mathbb{X}^*}-\overline{\mathbb{X}}}{S(\mathbb{X}^*)}$, où $\mathbb{X}^*$ est un échantillon bootstrap associé à $\mathbb{X}$.

Test bootstrap

Soit $\alpha \in]0,1[$. Le test bootstrap associé à la statistique $T(\mathbb{X})$ est donné par $\Phi_\alpha*(\mathbb{X}) = 1$ si $T(\mathbb{X}) > q^*_{1-\alpha}(\mathbb{X})$, 0 sinon, où $q^*_{1-\alpha}(\mathbb{X})$ est le $(1-\alpha)$ quantile de la loi conditionnelle de R_n^* sachant $\mathbb{X}$.

✔ Idem au test proposé initialement par Efron et Tibshirani (1993) (pp 226–227).
✔ Équivalence entre ce test et l'intervalle de confiance bootstrap-t.
✘ Si la loi de la racine $R_n(\mathbb{X}, P)$ semble très éloignée de la loi normale, on changera de statistique de test et de racine, typiquement $T(\mathbb{X}) = \overline{\mathbb{X}} - \theta_0$, et $R_n(\mathbb{X}, P) = \overline{\mathbb{X}} - \theta(P)$.

Tests bootstrap approchés

- Soit $\mathbb{X}^{*1}, \ldots, \mathbb{X}^{*B}$ B échantillons boostrap.
- Soit $R_n^{*b} = R_n(\mathbb{X}^{*b}, P_n)$.
- Soit $(R_n^{*(1)}, \ldots, R_n^{*(B)})$ la statistique d'ordre associée à $(R_n^{*1}, \ldots, R_n^{*B})$.

Le quantile conditionnel $q^*_{1-\alpha}(\mathbb{X})$ est approché par $R_n^{*(\lceil B - B\alpha \rceil)}$ (Monte Carlo).

- Premier test : on rejette (H_0) si $T(\mathbb{X}) > R_n^{*(\lceil B - B\alpha \rceil)}$.
- Deuxième test : étant donnée une observation x de $\mathbb{X}$, on rejette (H_0) si la p-valeur approchée $p^*_B(x) = \frac{\#\{b,\ r_n^{*b} \geq T(x)\}+1}{B+1}$ est inférieure ou égale à α.

Algorithme Test bootstrap sur la moyenne

Variable

B : 999, 9999... *tel que* $(B+1)\alpha$ *entier*

Début

Pour b **variant de** 1 **à** B

Générer x^{*b} réalisation d'un échantillon bootstrap

Calculer $r_n^{*b} = R_n(x^{*b}, P_n)$ réplique bootstrap de $R_n(x, P)$

FinPour

Retourner $p_B^*(x) = \frac{\#\{b,\ r_n^{*b} \geq T(x)\}+1}{B+1}$

Fin

Le coin du UseR

```
myT=function(data,index){
  sqrt(length(data))*(mean(data[index])-mean(data))/sd(data[index]) };
pval=function(X,B,theta0){
  Tbst=boot(X,myT,R=B);
  Tnull=sqrt(length(X))*(mean(X)-theta0)/sd(X);
  return((1+sum(Tbst$t>=Tnull))/(B+1)) };
```

$\hookrightarrow$ Généralisations :

- Test bilatère (redéfinir le test approché à l'aide la p-valeur),
- Test sur un paramètre d'intérêt $\theta(P)$ dont on dispose d'un "bon" estimateur $\hat{\theta} = T(\mathbb{X})$.

Illustration

Test de (H_0) $\theta(P) = 1$ versus (H_1) $\theta(P) \neq 1$.
Estimation de la puissance du test pour $\mathbb{X} = 30$-échantillon d'une loi exponentielle de différents paramètres.

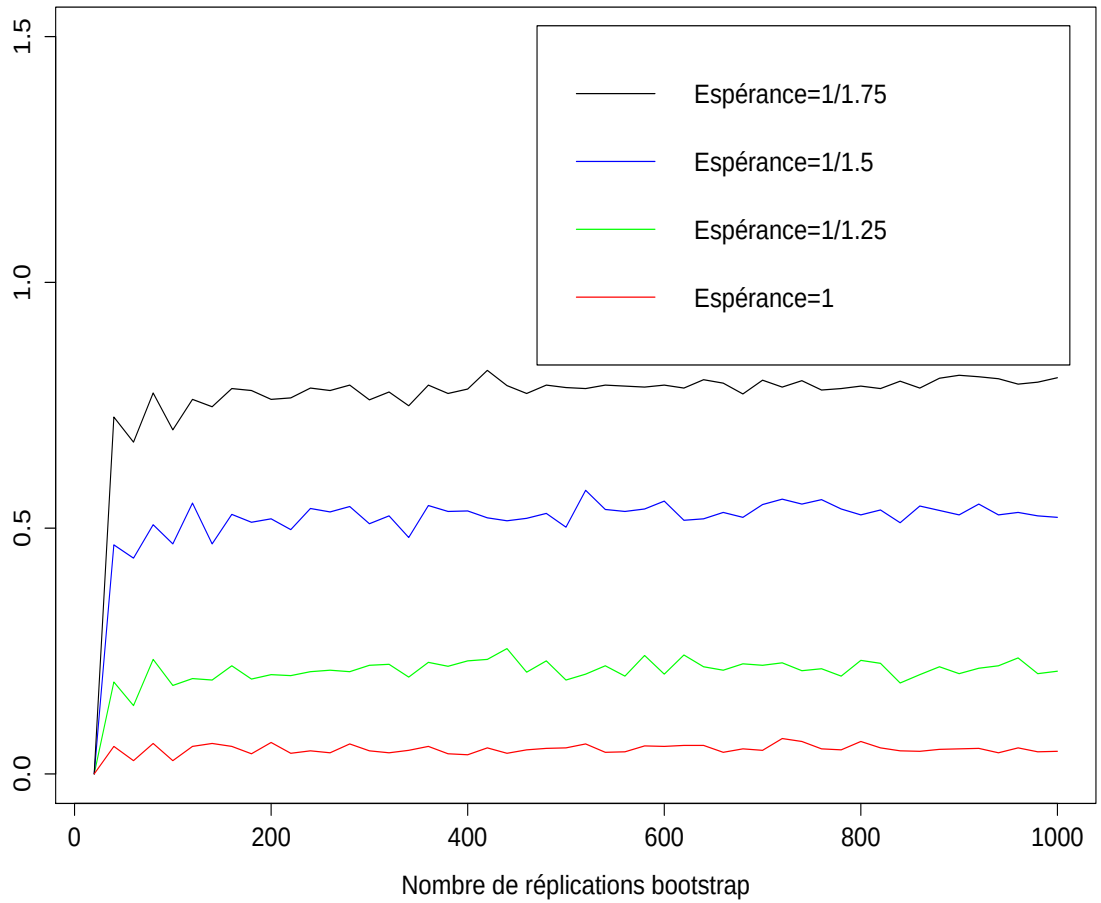

Figure: Puissance du test bootstrap de la moyenne

Tests d'hypothèses

Éléments de bibliographie

Ouvrages

- Fisher, R. A. (1935). The Design of Experiments.
- Good, P. (2004). Permutation, Parametric, and Bootstrap Tests of Hypotheses.
- Van der Vaart, A., and Wellner, J. A. (1996). Weak convergence and empirical processes. Pages 360–366, pour le test de Kolmogorov-Smirnov généralisé.

Articles

- Hoeffding, W. (1952). The large-power of tests based on permutations of observations.
- Romano, J. P. and Wolf, M. (2005). Exact and Approximate Stepdown Methods for Multiple Hypothesis Testing.