Chapitre 3 :
Méthode du bootstrap

- Introduction
- Le théorème de GlivenkoCantelli
- Bootstrap
- Bootstrap paramétrique

Aléa intrinsèque

Estimation à partir d'un échantillon aléatoire = incertitude

Puisque fondé sur un échantillon **aléatoire**, un estimateur

$$\delta(X_1, \ldots, X_n)$$

est aussi (une variable) **aléatoire**

Variation moyenne

Question 1 :

De combien varie $\delta(X_1, \ldots, X_n)$ quand l'échantillon varie ?

Question 2 :

Quelle est la variance de $\delta(X_1, \ldots, X_n)$?

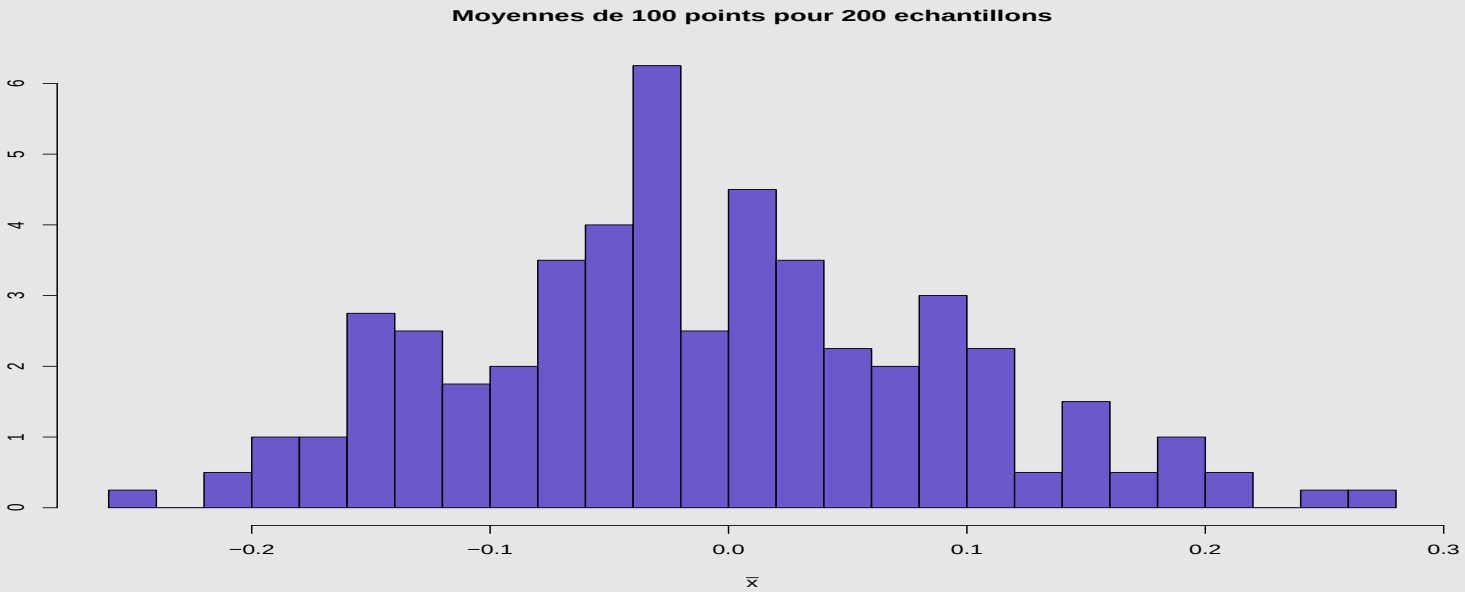
Question 3 :

Quelle est la distribution de $\delta(X_1, \ldots, X_n)$?

Example (**Échantillon normal)**

Soit $X_1, \ldots, X_{100}$ un échantillon normal $\mathcal{N}(\theta, 1)$. Sa moyenne θ est estimée par

$$\hat{\theta} = \frac{1}{100} \sum_{i=1}^{100} X_i$$



Variation compatible avec la loi (connue) $\hat{\theta} \sim \mathcal{N}(\theta, 1/100)$

Problèmes correspondants

- On observe **un seul** échantillon en général
- La loi de l'échantillon est souvent inconnue
- L'évaluation de la variation moyenne de $\delta(X_1, \ldots, X_n)$ est essentielle pour la construction d'intervalles de confiance et de tests de/réponses à des questions comme

$$\mathrm{H}_0 \ : \ \theta \leq 0$$

- En cas de **normalité** de l'échantillon, le **vrai** θ se trouve avec forte probabilité dans l'intervalle

$$[\hat{\theta} - 2\sigma, \hat{\theta} + 2\sigma] \, .$$

Quid de σ ?!

Estimation de la fonction de répartition

Extension/application de la LGN à l'approximation de la fonction de répartition :
Pour un échantillon $X_1, \ldots, X_n$, si

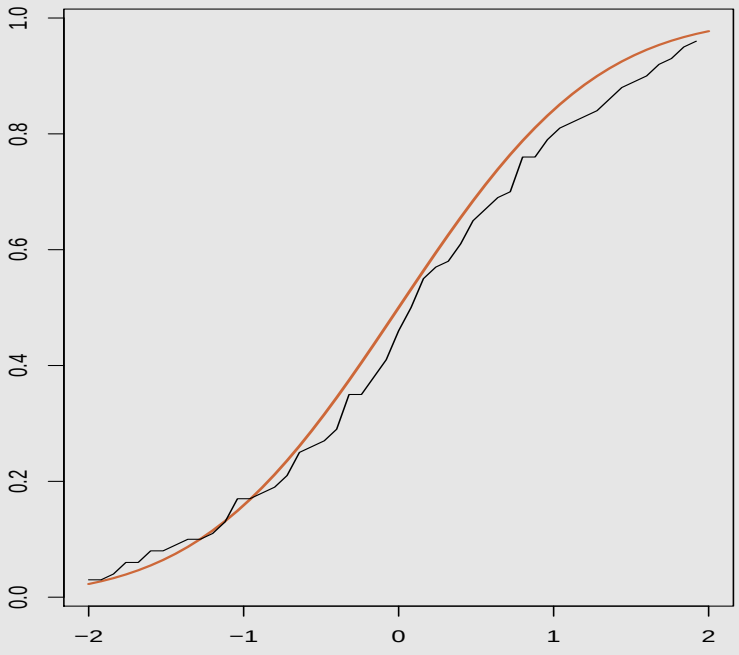
$$\begin{aligned} \hat{F}_n(x) &= \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{]-\infty, X_i]}(x) \\ &= \frac{\text{card}\left\{X_i;\ X_i \leq x\right\}}{n}, \end{aligned}$$

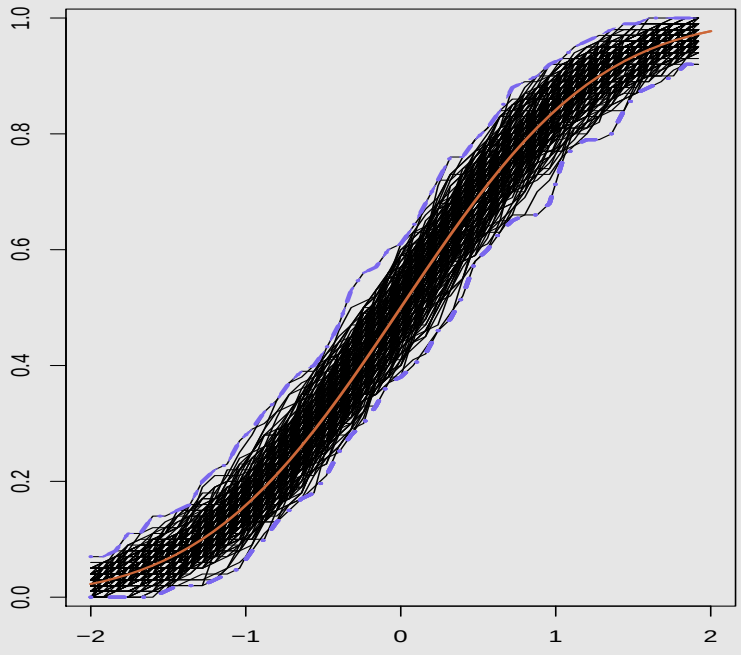
$\hat{F}_n(x)$ est un estimateur convergent de la fonction de répartition $F(x)$

[Glivenko–Cantelli]

$$\hat{F}_n(x) \longrightarrow \text{Pr}(X \leq x)$$

Example (**Échantillon normal**)





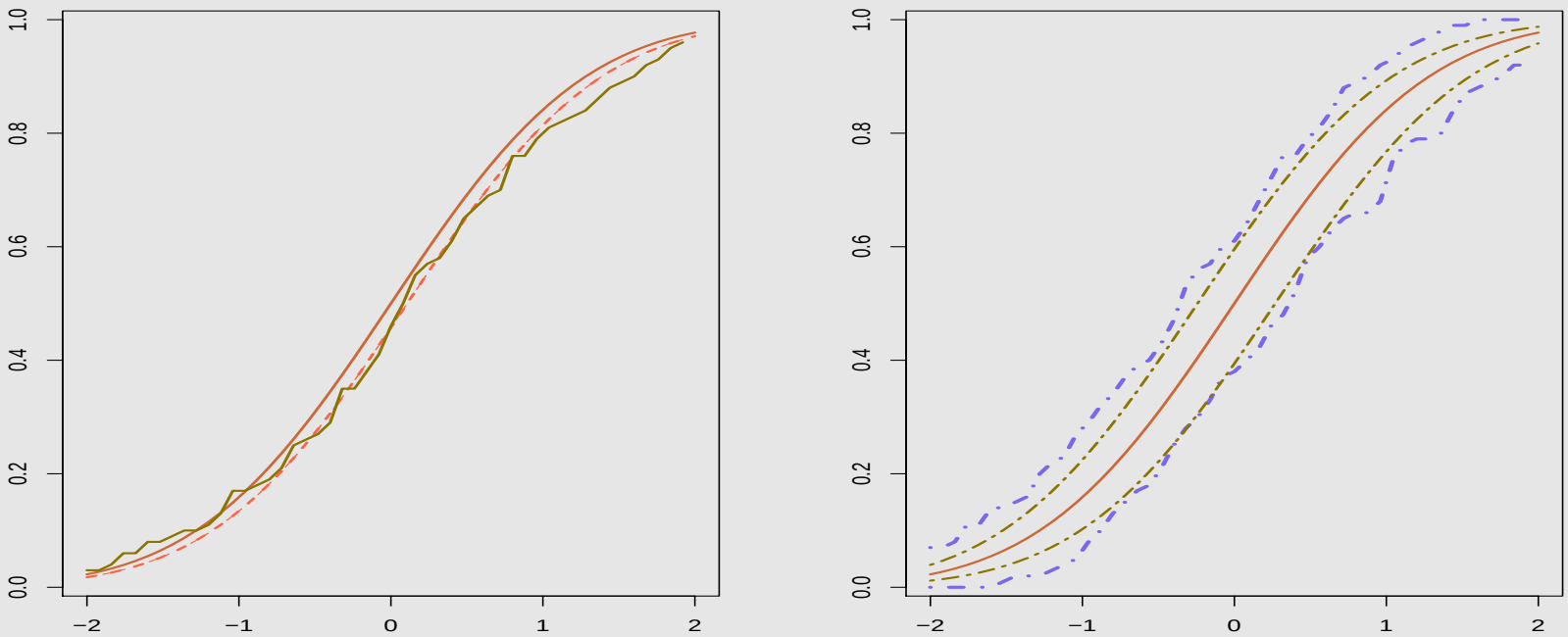
Estimation de la fonction de répartition F à partir d'un échantillon normal de 100 points et variation de cette estimation sur 200 échantillons normaux

Propriétés

- Estimateur dit *non-paramétrique* : on n'a pas besoin de la loi ni de la forme de la loi de l'échantillon pour construire cet estimateur © **Il est toujours disponible**
- **Robustesse contre efficacité :** si la forme [paramétrique] de la loi est connue, meilleure approximation fondée sur cette forme, mais si on se trompe de [forme de] loi, le résultat peut être bien pire !

Example (**Échantillon normal**)

Fonction de répartition de $\mathcal{N}(\theta, 1)$, $\Phi(x - \theta)$



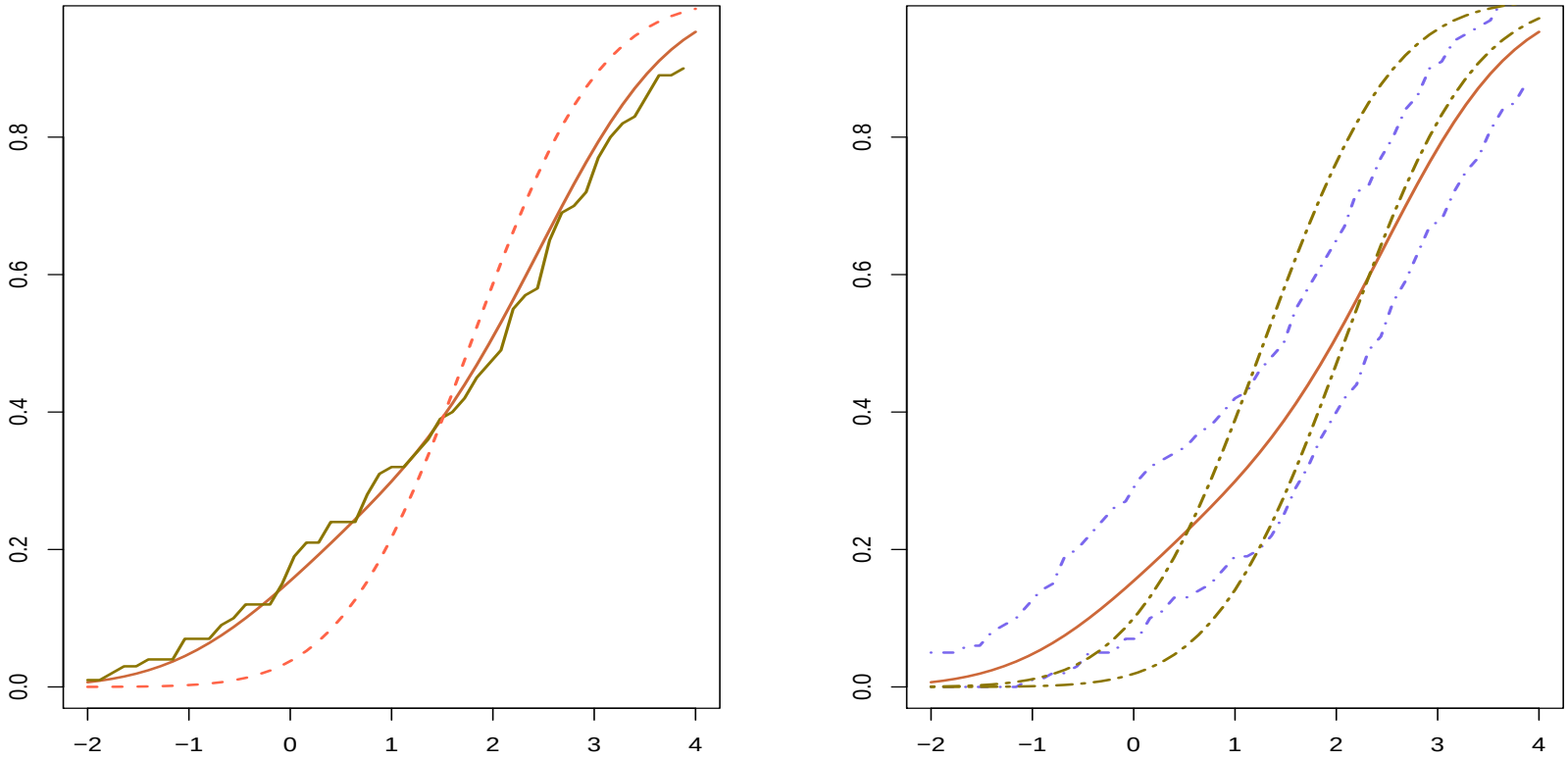
Estimation de $\Phi(\cdot - \theta)$ par $\hat{F}_n$ et $\Phi(\cdot - \hat{\theta})$ à partir de 100 points et variation maximale de ces estimations sur 200 réplications

Example (**Échantillon non-normal)**

Echantillon provenant de

$$0.3\mathcal{N}(0,1) + 0.7\mathcal{N}(2.5,1)$$

faussement alloué à une loi normale $\Phi(\cdot - \theta)$



Estimation de F par $\hat{F}_n$ et $\Phi(\cdot - \hat{\theta})$ à partir d'un échantillon de mélange de 100 points et variation de ces estimations sur 200 échantillons de mélange

Extension aux fonctionnelles de F

Pour toute expression de la forme

$$\theta(F) = \int h(x)\, dF(x)\,,$$

[Fonctionnelle de la cdf]

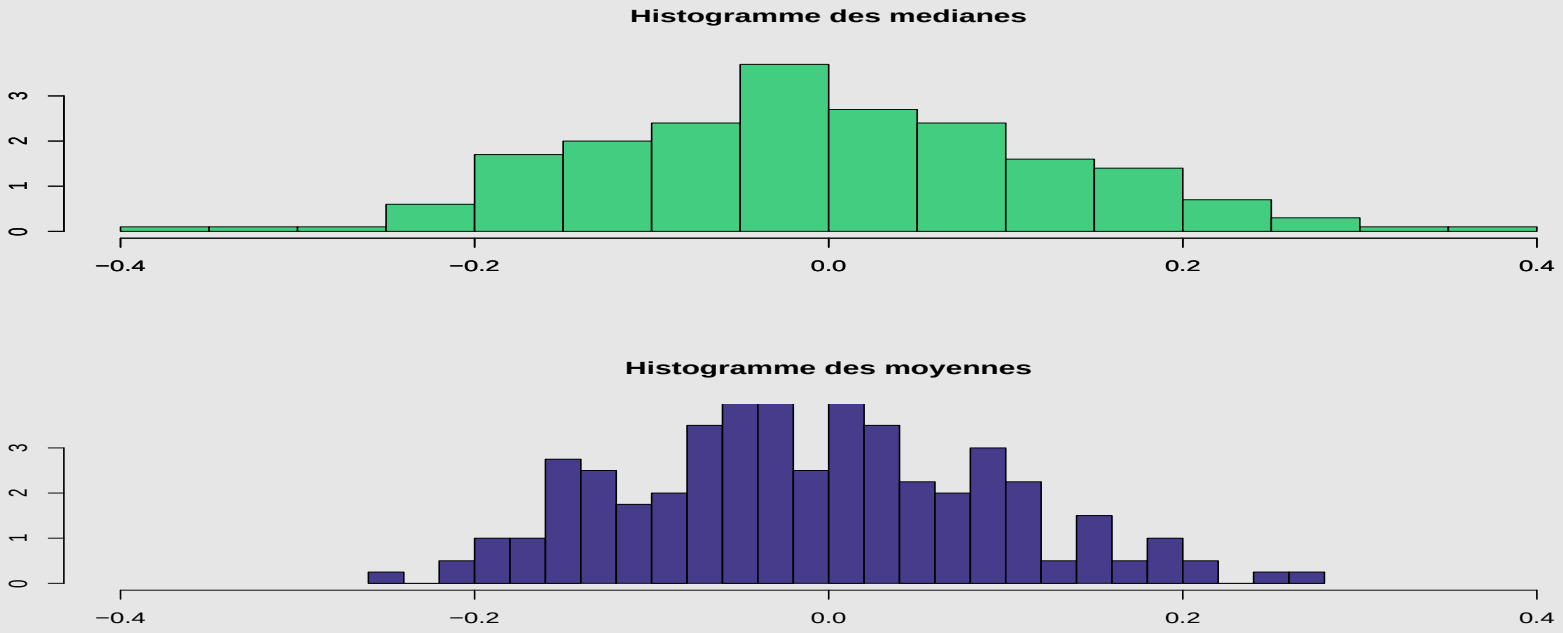
utilisation de l'approximation

$$\begin{aligned} \widehat{\theta(F)} &= \theta(\hat{F}_n) \\ &= \int h(x)\, d\hat{F}_n(x) \\ &= \frac{1}{n}\sum_{i=1}^{n} h(X_i) \end{aligned}$$

[Estimateur des moments]

Example (**Échantillon normal**)

Comme θ est (aussi) la médiane de $\mathcal{N}(\theta, 1)$, $\hat{\theta}$ peut être pris comme médiane de $\hat{F}_n$, donc comme médiane de $X_1, \ldots, X_n$, soit $X_{(n/2)}$



Comparaison des variations des moyennes et des médianes sur 200 échantillons normaux

Comment approcher la distribution de $\theta(\hat{F}_n)$?

Principe

Comme

$$\theta(\hat{F}_n) = \theta(X_1, \ldots, X_n) \quad \textbf{avec} \quad X_1, \ldots, X_n \overset{i.i.d.}{\sim} F$$

on remplace F par $\hat{F}_n$:

$$\theta(\hat{F}_n) \approx \theta(X_1^*, \ldots, X_n^*) \quad \textbf{avec} \quad X_1^*, \ldots, X_n^* \overset{i.i.d.}{\sim} \hat{F}_n$$

Implémentation

$\hat{F}_n$ étant connue, on peut simuler suivant $\hat{F}_n$, donc approcher la loi de $\theta(X_1^*, \ldots, X_n^*)$ [au lieu de celle de $\theta(X_1, \ldots, X_n)$]
La loi correspondant à

$$\hat{F}_n(x) = \frac{\text{card}\,\{X_i;\ X_i \leq x\}}{n}$$

donne une probabilité de $1/n$ à chaque point de $\{x_1, \ldots, x_n\}$:

$$\Pr^{\hat{F}_n}(X^* = x_i) = \frac{1}{n}$$

Il suffit donc d'opérer des tirages **avec remise** dans $(X_1, \ldots, X_n)$

[en R, sample(x,n,replace=T)]

Simulation par Monte Carlo

1. Pour $b = 1, \ldots, B$,
   1. générer un échantillon $X_1^b, \ldots, X_n^b$ suivant $\hat{F}_n$
   2. construire l'image correspondante

$$\hat{\theta}^b = \theta(X_1^b, \ldots, X_n^b)$$

2. Utiliser l'échantillon

$$\hat{\theta}^1, \ldots, \hat{\theta}^B$$

pour approcher la distribution de

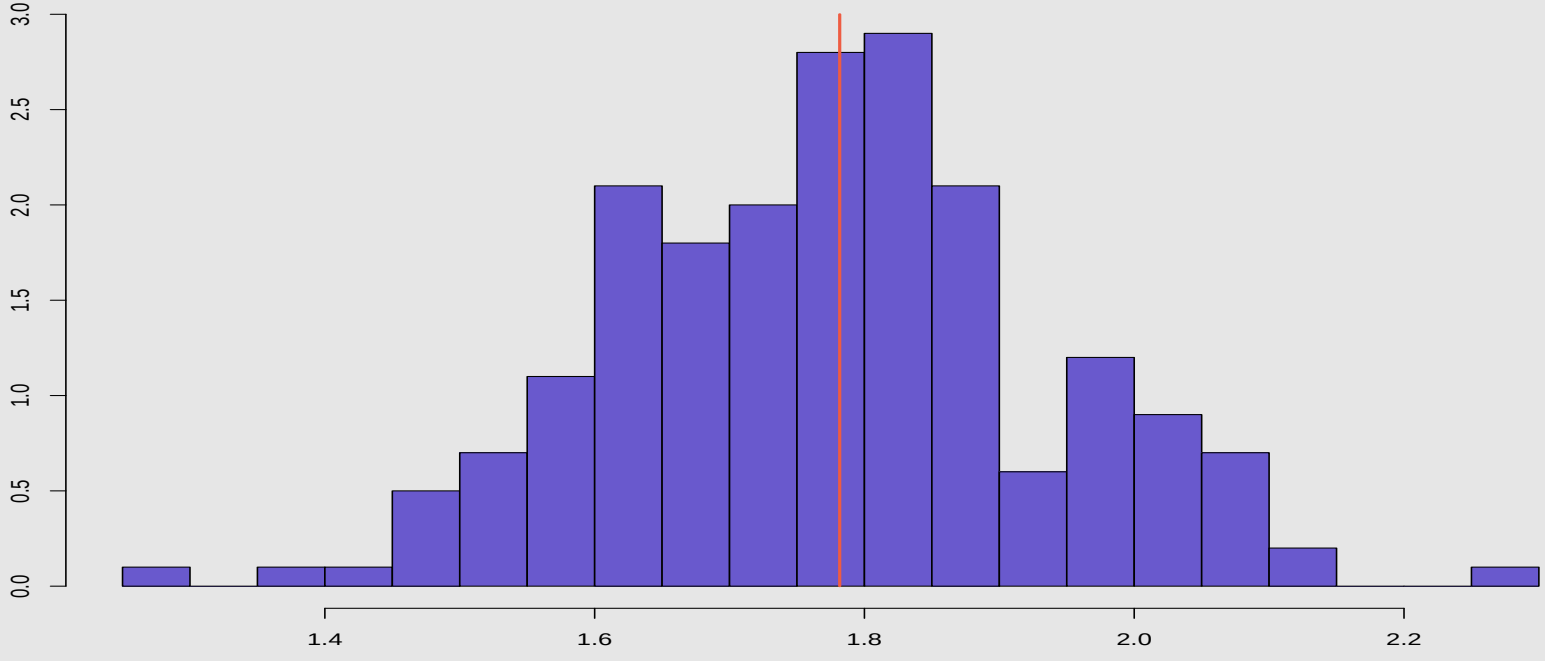
$$\theta(X_1, \ldots, X_n)$$

Notes

- `bootstrap = languette de botte`
  on utilise seulement l'échantillon pour construire une évaluation de sa loi

  [Aventures du Baron de Munchausen]
- un échantillon bootstrap est obtenu par n tirages avec remise dans $(X_1, \ldots, X_n)$
- il peut donc prendre n^n valeurs (ou $\binom{2n-1}{n}$ valeurs si on ne considère pas l'ordre)

Example (**Échantillon** $0.3\mathcal{N}(0,1) + 0.7\mathcal{N}(2.5,1)$)



Variation des moyennes empiriques sur 200 échantillons bootstrap et moyenne de l'échantillon observé

Example (**Calcul de la variation moyenne**)

Pour un estimateur $\theta(X_1, \ldots, X_n)$, l'écart-type est donné par

$$\eta(F) = \sqrt{\mathsf{E}^F[(\theta(X_1, \ldots, X_n) - \mathsf{E}^F[\theta(X_1, \ldots, X_n)])^2]}$$

et son approximation bootstrap est

$$\eta(\hat{F}_n) = \sqrt{\mathsf{E}^{\hat{F}_n}[(\theta(X_1, \ldots, X_n) - \mathsf{E}^{\hat{F}_n}[\theta(X_1, \ldots, X_n)])^2]}$$

Example (**Calcul de la variation moyenne (2)**)

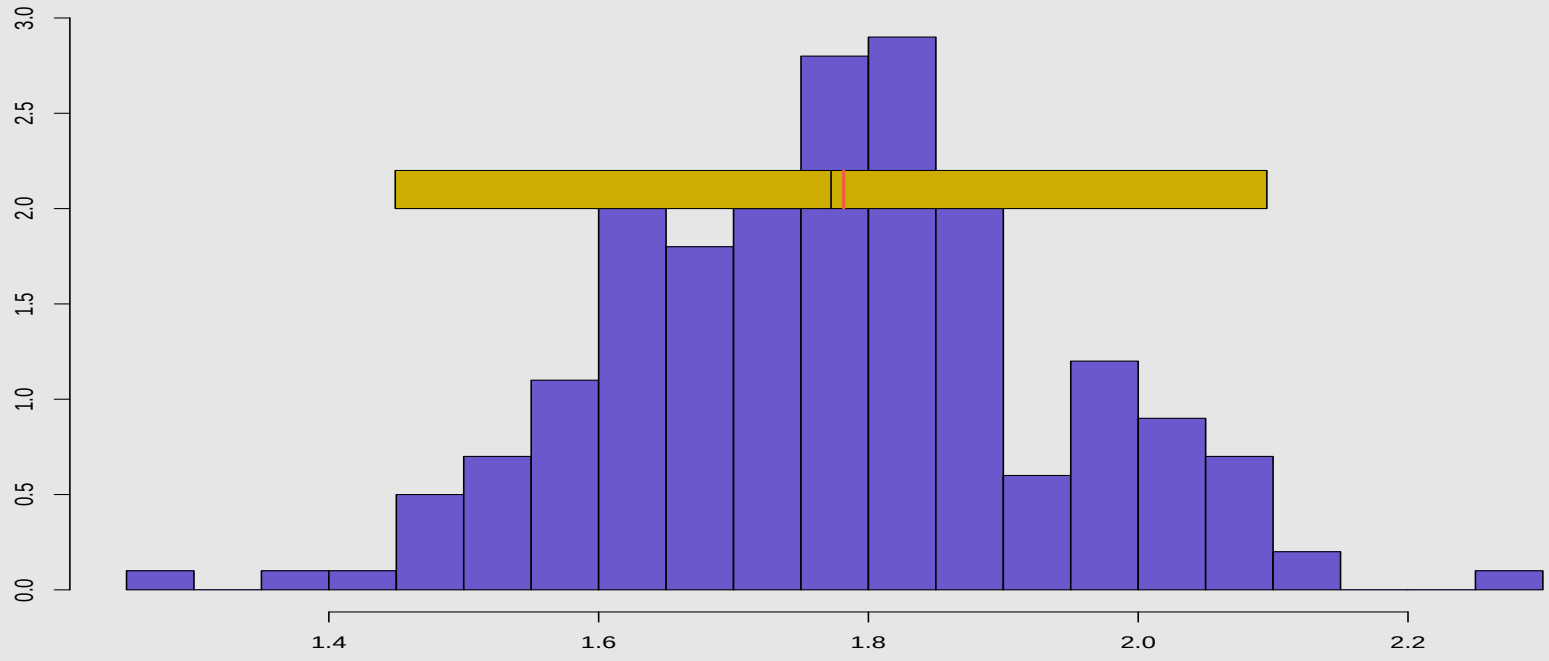
Approximation elle-même approchée par

$$\hat{\eta}(\hat{F}_n) = \left(\frac{1}{B} \sum_{b=1}^{B} (\theta(X_1^b, \ldots, X_n^b) - \bar{\theta})^2 \right)^{1/2}$$

où

$$\bar{\theta} = \frac{1}{B} \sum_{b=1}^{B} \theta(X_1^b, \ldots, X_n^b)$$

Example (**Échantillon** $0.3\mathcal{N}(0,1) + 0.7\mathcal{N}(2.5,1)$**)**



Intervalle de variation bootstrap à $\pm 2\hat{\eta}(\hat{F}_n)$ et moyenne de l'échantillon observé

Example (**Échantillon normal**)

Echantillon

$$(X_1, \ldots, X_{100}) \overset{i.i.d.}{\sim} \mathcal{N}(\theta, 1)$$

Comparaison des intervalles de confiance

$$[\bar{x} - 2 * \hat{\sigma}_x/10, \bar{x} + 2 * \hat{\sigma}_x/10] = [-0.113, 0.327]$$

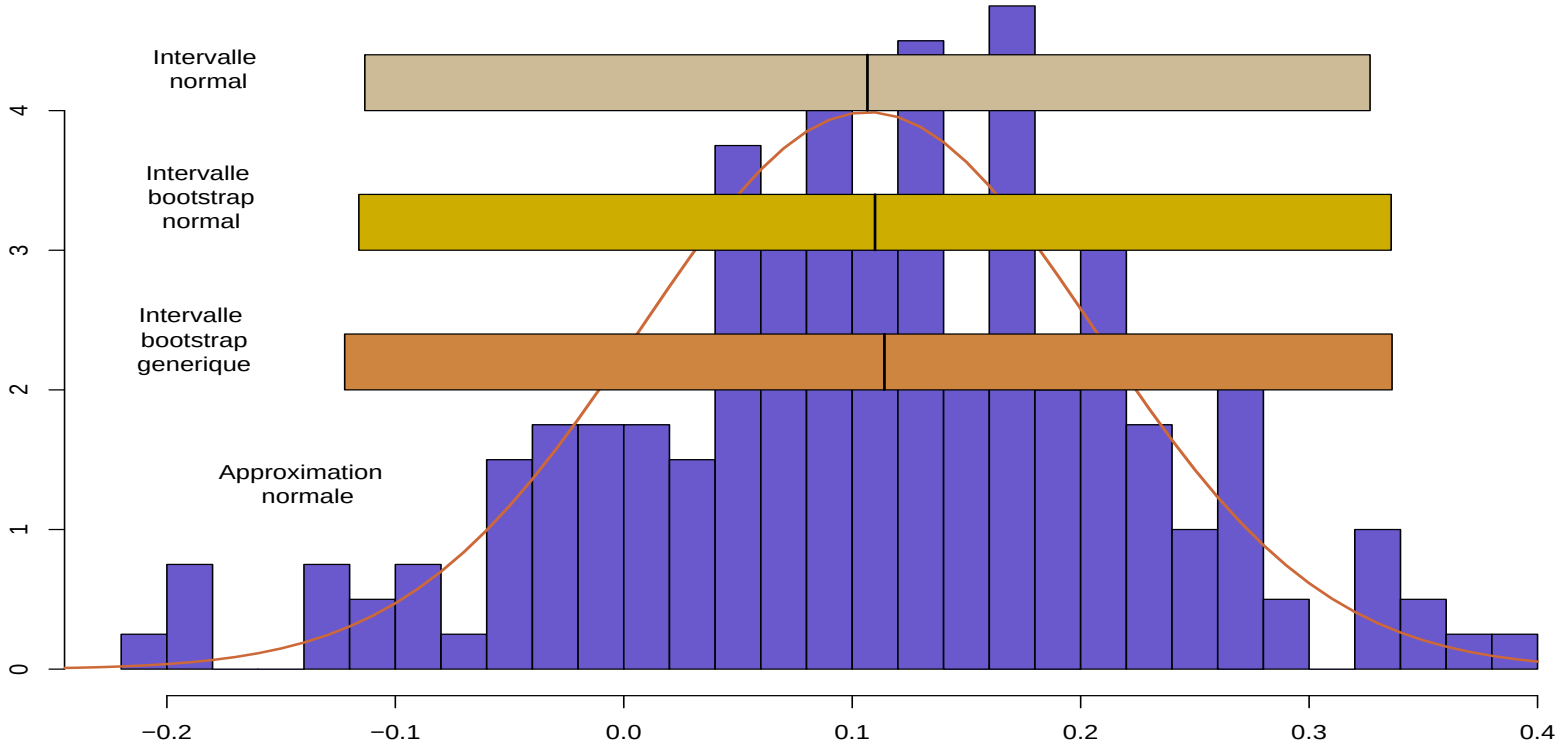
[approximation normale]

$$[\bar{x}^* - 2 * \hat{\sigma}^*, \bar{x}^* + 2 * \hat{\sigma}^*] = [-0.116, 0.336]$$

[approximation bootstrap normale]

$$[q^*(0.025), q^*(0.975)] = [-0.112, 0.336]$$

[approximation bootstrap générique]



Intervalles de variation à 95% pour un échantillon de 100 points et 200 répliques bootstrap

Boostrap paramétré

Si la forme paramétrique de F est connue,

$$F(\cdot) = \Phi_\lambda(\cdot) \qquad \lambda \in \Lambda\,,$$

une évaluation de F plus efficace que $\hat{F}_n$ est fournie par

$$\Phi_{\hat{\lambda}_n}$$

où $\hat{\lambda}_n$ est un estimateur convergent de λ

[Cf Exemple 52]

Boostrap paramétrique

Approximation de la loi de

$$\theta(X_1, \ldots, X_n)$$

par la loi de

$$\theta(X_1^*, \ldots, X_n^*) \qquad X_1^*, \ldots, X_n^* \overset{i.i.d.}{\sim} \Phi_{\hat{\lambda}_n}$$

Peut éviter le recours à la simulation dans certains cas

Example (**Échantillon exponentiel**)

Soit

$$X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathcal{E}\mathrm{xp}(\lambda)$$

et $\lambda = 1/\mathrm{E}_\lambda[X]$ à estimer.
Un estimateur possible est

$$\hat{\lambda}(x_1, \ldots, x_n) = \frac{n}{\sum_{i=1}^n x_i}$$

mais cet estimateur est biaisé :

$$\mathrm{E}_\lambda[\hat{\lambda}(X_1, \ldots, X_n)] \neq \lambda$$

Example (**Échantillon exponentiel (2)**)

Questions :

- Comment évaluer le biais

$$\lambda - \mathsf{E}_\lambda[\hat{\lambda}(X_1, \ldots, X_n)]$$

  de cet estimateur ?
- Quelle est la loi de cet estimateur ?

Evaluation bootstrap du biais

Example (**Échantillon exponentiel (3)**)

$$\hat{\lambda}(x_1, \ldots, x_n) - \mathsf{E}_{\hat{\lambda}(x_1,\ldots,x_n)}[\hat{\lambda}(X_1, \ldots, X_n)]$$

[Forme paramétrique]

$$\hat{\lambda}(x_1, \ldots, x_n) - \mathsf{E}_{\hat{F}_n}[\hat{\lambda}(X_1, \ldots, X_n)]$$

[Forme non-paramétrique]

Example (**Échantillon exponentiel (4)**)

Dans le premier cas (paramétrique),

$$1/\hat{\lambda}(X_1, \ldots, X_n) \sim \mathcal{G}a(n, n\lambda)$$

et

$$\mathsf{E}_\lambda[\hat{\lambda}(X_1, \ldots, X_n)] = \frac{n}{n-1}\lambda$$

donc le biais est **analytiquement** évalué comme

$$-\lambda/n - 1$$

estimé par

$$-\frac{\hat{\lambda}(X_1, \ldots, X_n)}{n-1} = -0.00787$$

Example (Échantillon exponentiel (5))

Dans le second cas (non-paramétrique), évaluation par Monte Carlo,

$$\hat{\lambda}(x_1, \ldots, x_n) - \mathsf{E}_{\hat{F}_n}[\hat{\lambda}(X_1, \ldots, X_n)] = 0.00142$$

qui est du **"mauvais"** signe

Example (**Échantillon exponentiel (6)**)

Construction d'un intervalle de confiance sur λ

Par bootstrap paramétrique,

$$\Pr_\lambda\left(\hat{\lambda}_1 \leq \lambda \leq \hat{\lambda}_2\right) = \Pr\left(\omega_1 \leq \lambda/\hat{\lambda} \leq \omega_2\right) = 0.95$$

peut être déduit de

$$\lambda/\hat{\lambda} \sim \mathcal{G}a(n,n)$$

[En R, qgamma(0.975,n,1/n)]

$$[\hat{\lambda}_1, \hat{\lambda}_2] = [0.452, 0.580]$$

Example (**Échantillon exponentiel (7)**)

Par bootstrap non-paramétrique, on remplace

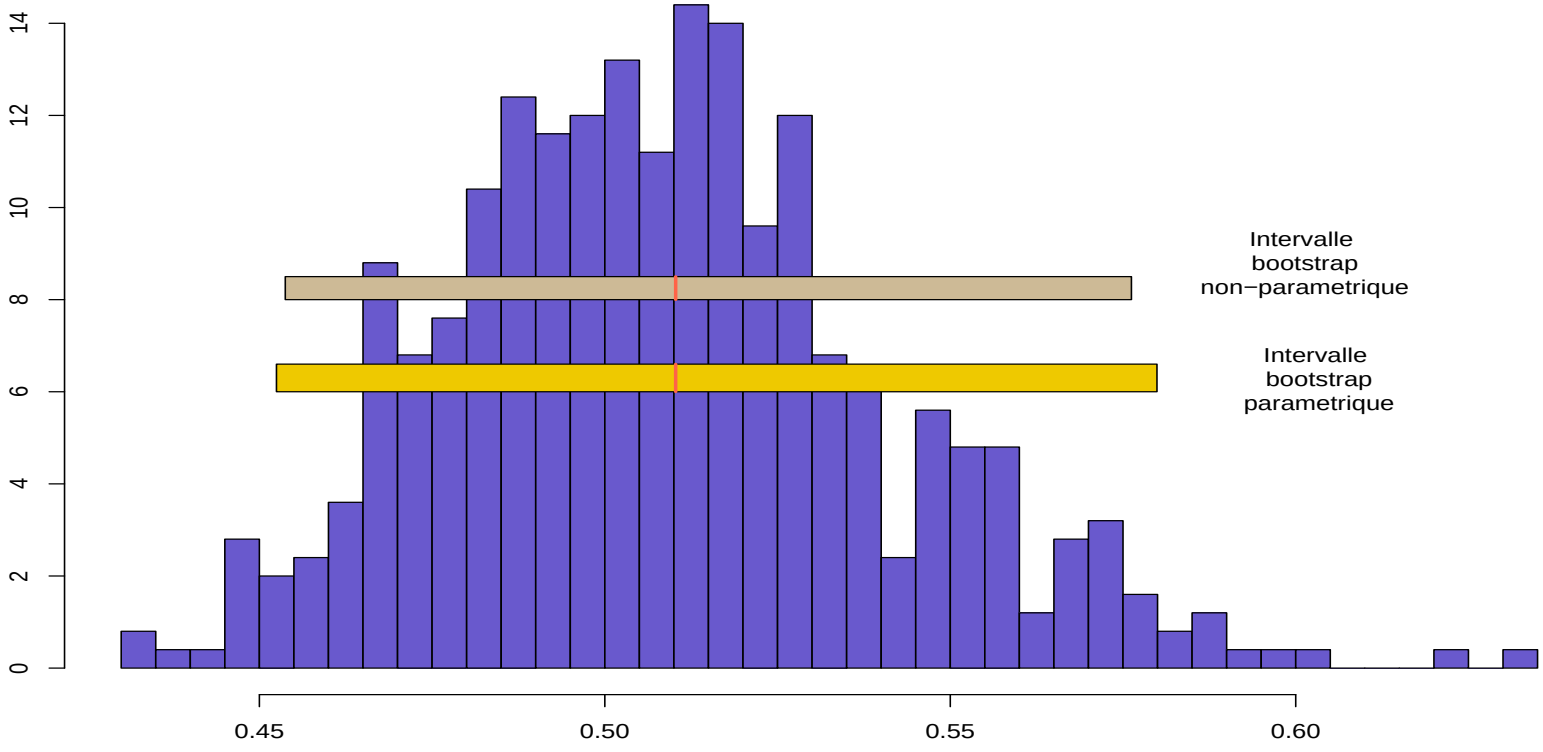
$$\Pr_F \left(q(.025) \leq \lambda(F) \leq q(.975)\right) = 0.95$$

par

$$\Pr_{\hat{F}_n} \left(q^*(.025) \leq \lambda(\hat{F}_n) \leq q^*(.975)\right) = 0.95$$

Approximation des quantiles $q^*(.025)$ et $q^*(.975)$ de $\lambda(\hat{F}_n)$ par échantillonnage bootstrap (Monte Carlo)

$$[q^*(.025), q^*(.975)] = [0.454, 0.576]$$


0
2
4
6
8
10
12
14
0.45
0.50
0.55
0.60
Intervalle
bootstrap
non−parametrique
Intervalle
bootstrap
parametrique

Example (**Échantillon Student**)

Soit

$$X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathfrak{T}(5, \mu, \tau^2) \overset{\text{def}}{=} \mu + \tau \frac{\mathcal{N}(0,1)}{\sqrt{\chi^2_5/5}}$$

On peut alors estimer μ et τ par

$$\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \begin{aligned} \hat{\tau}_n &= \sqrt{\frac{5-2}{5}}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu})^2} \\ &= \sqrt{\frac{5-2}{5}}\,\hat{\sigma}_n \end{aligned}$$

Example (**Échantillon Student (2)**)

Problème

$\hat{\mu}_n$ n'est pas distribuée comme une loi de Student $\mathfrak{T}(5, \mu, \tau^2/n)$
On doit donc reconstituer la loi de $\hat{\mu}_n$ par échantillonnage bootstrap.

Example (**Échantillon Student (3)**)

Comparaison des intervalles de confiance

$$[\hat{\mu}_n - 2 * \hat{\sigma}_n/10, \hat{\mu}_n + 2 * \hat{\sigma}_n/10] = [-0.068, 0.319]$$

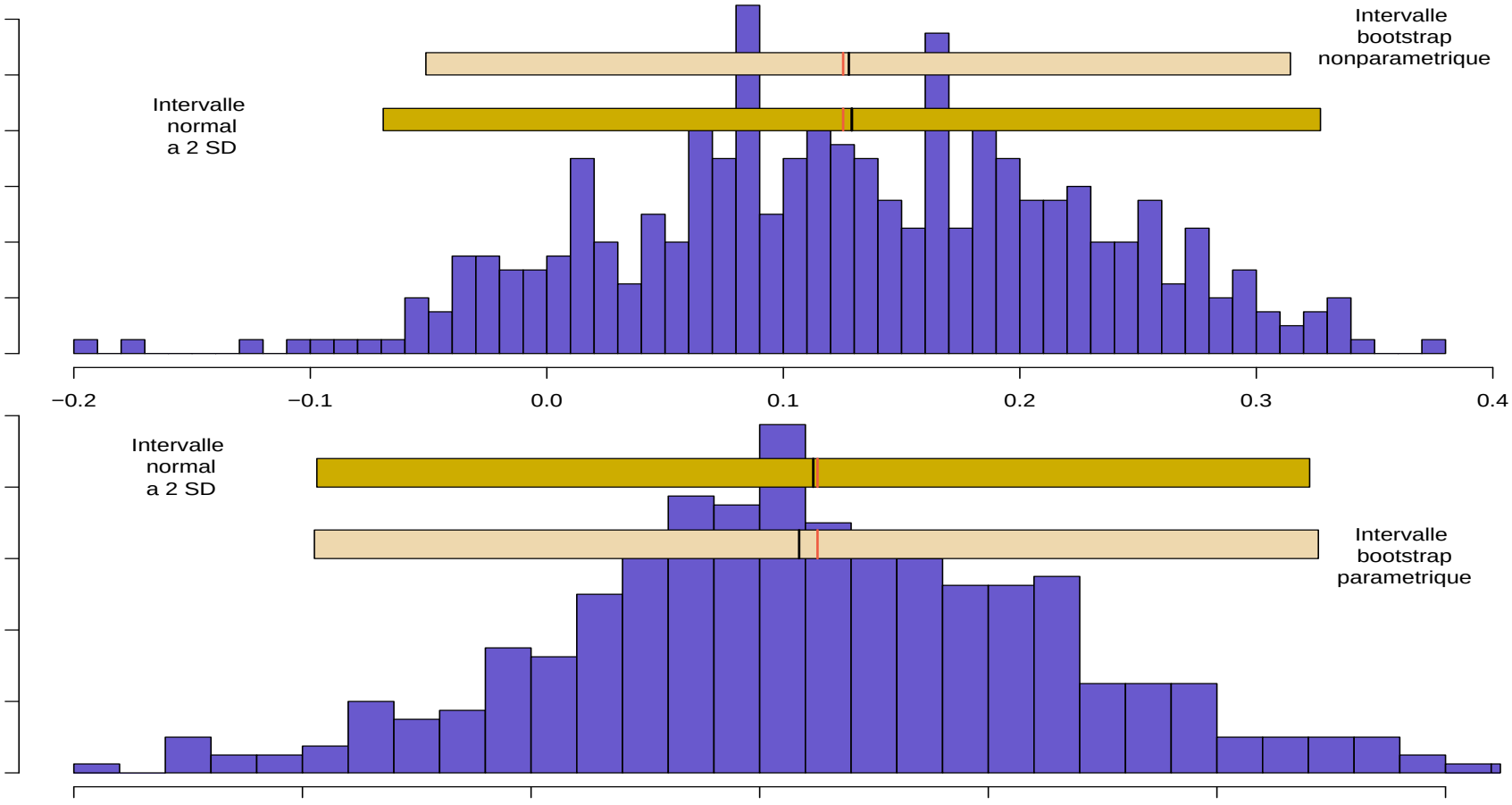
[approximation normale]

$$[q^*(0.05), q^*(0.95)] = [-0.056, 0.305]$$

[approximation bootstrap paramétrique]

$$[q^*(0.05), q^*(0.95)] = [-0.094, 0.344]$$

[approximation bootstrap non-paramétrique]



Intervalles de variation à 95% pour un échantillon de 150 points et 400 répliques bootstrap (haut) non-paramétriques et (bas) paramétriques