UNE INTRODUCTION AU BOOTSTRAP

Bernard Rapacchi
Centre Interuniversitaire de Calcul de Grenoble

15 décembre 1994

Table des matières

0.1 Une petite introduction 1
0.1.1 La précision d'une moyenne 1
0.1.2 L'idée du bootstrap 2
0.2 Quelques rappels 4
0.2.1 Echantillons aléatoires 4
0.2.2 probabilités et statistiques 5
0.2.3 L'exemple des écoles 6
0.3 Estimateur "Plug-in" 7
0.3.1 Fonction de distribution empirique 7
0.3.2 Comment sont liés y et z? 8
0.3.3 Le principe "je branche" 9
0.4 Erreurs standards 10
0.4.1 Erreur-standard d'une moyenne 10
0.4.2 Théorème central limite 11
0.4.3 Tirages à pile ou face 12
0.5 Estimateurs bootstrap de l'erreur standard 13
0.5.1 Estimateur bootstrap non-papramètrique de l'erreur standard 13
0.5.2 Algorithme d'estimation des erreurs-standards 14
0.5.3 Algorithme pour l'estimateur bootstrap de l'erreur-standard 15
0.5.4 un exemple 16
0.5.5 Combien de bootstraps? 17

0.5.6 Estimateur bootstrap paramétrique de l'erreur-standard 18
0.5.7 Avant le bootstrap 19
0.6 Exemple d'utilisation 20
0.6.1 L'analyse factorielle 20
0.6.2 Pourquoi une acp? 21
0.6.3 Le bootstrap 22
0.6.4 Et les vecteurs propres 25
0.7 Structures de données plus complexes 27
0.7.1 Généralités 27
0.7.2 Deux échantillons 28
0.7.3 Bootstrap 29
0.7.4 Structures de données plus générales . . . 31
0.7.5 Exemple: série chronologique 32
0.7.6 Comment utiliser le bootstrap? 33
0.8 Régression linéaire 36
0.8.1 Présentation du problème 36
0.8.2 modèle probabiliste 37
0.8.3 Bootstrap 38
0.8.4 bootstrap par paires ou par résidus? . . . 40
0.9 Estimation du biais 41
0.9.1 Présentation du problème 41
0.9.2 Exemple: les patchs 42
0.9.3 Bootstrap 43
0.9.4 Loi des grands nombres 45
0.10 Le Jacknife 46
0.10.1 Présentation 46
0.10.2 Estimation Jacknife du biais 47

0.10.3 Exemple : tests étudiants 48
0.10.4 Relation bootstrap et jacknife 49
0.10.5 Problèmes du jacknife 50
0.10.6 D-jacknife 51
0.11 Intervalles de confiance et Bootstrap 52
0.11.1 Intervalle "normalisé" 52
0.11.2 Intervalle "t-Studentisé" 54
0.11.3 Intervalle "t-bootstrapé" 55
0.11.4 Exemple : souris 56
0.11.5 Transformations 57
0.11.6 Intervalle des percentiles 58
0.11.7 Intervalle accéléré et non-biaisé 60
0.12 Tests de permutation 62
0.12.1 Historique 62
0.12.2 Exemple : les souris 63
0.12.3 L'idée 64
0.12.4 Calcul de la statistique par test de permutation 66
0.12.5 Un autre exemple 68

0.1 Une petite introduction

0.1.1 La précision d'une moyenne

LA PRÉCISION D'UNE MOYENNE

Les Souris.

Temps de survie en jours après une intervention chirurgicale.

Groupe	Données	Taille	Moyenne	Erreur-standard
Traitées	94 197 16 38 99 141 23	7	86.86	25.24
Contrôle	52 104 146 10 51 30 40 27 46	9	56.22	14.14
		Différence:	30.63	28.93

$$\overline{x} = \sum_{i=1}^{n} x_i/n$$

$$\text{erreur-standard} = \sqrt{s^2/n}$$

$$s^2 = \sum_{i=1}^{n} (x_i - \overline{x})/(n-1)$$

Qu'est ce qui ce passe si on veut comparer les 2 groupes par leur médiane?

0.1.2 L'idée du bootstrap

L'IDÉE du BOOTSTRAP ou du CYRANO

– n iid $\mathbf{x} = (x_1, x_2, \ldots, x_n)$

– on est intéressé par une statistique $s(\mathbf{x})$

– on tire B échantillons $\mathbf{x}_b^* = (x_1^*, x_2^*, \ldots, x_n^*)$ où chaque $\mathbf{x}^*$ est constitué en tirant avec remise n valeurs parmi les x_i.

$$\text{exemple: } \mathbf{x}^* = (x_7, x_1, x_7, x_3, \ldots, x_4)$$

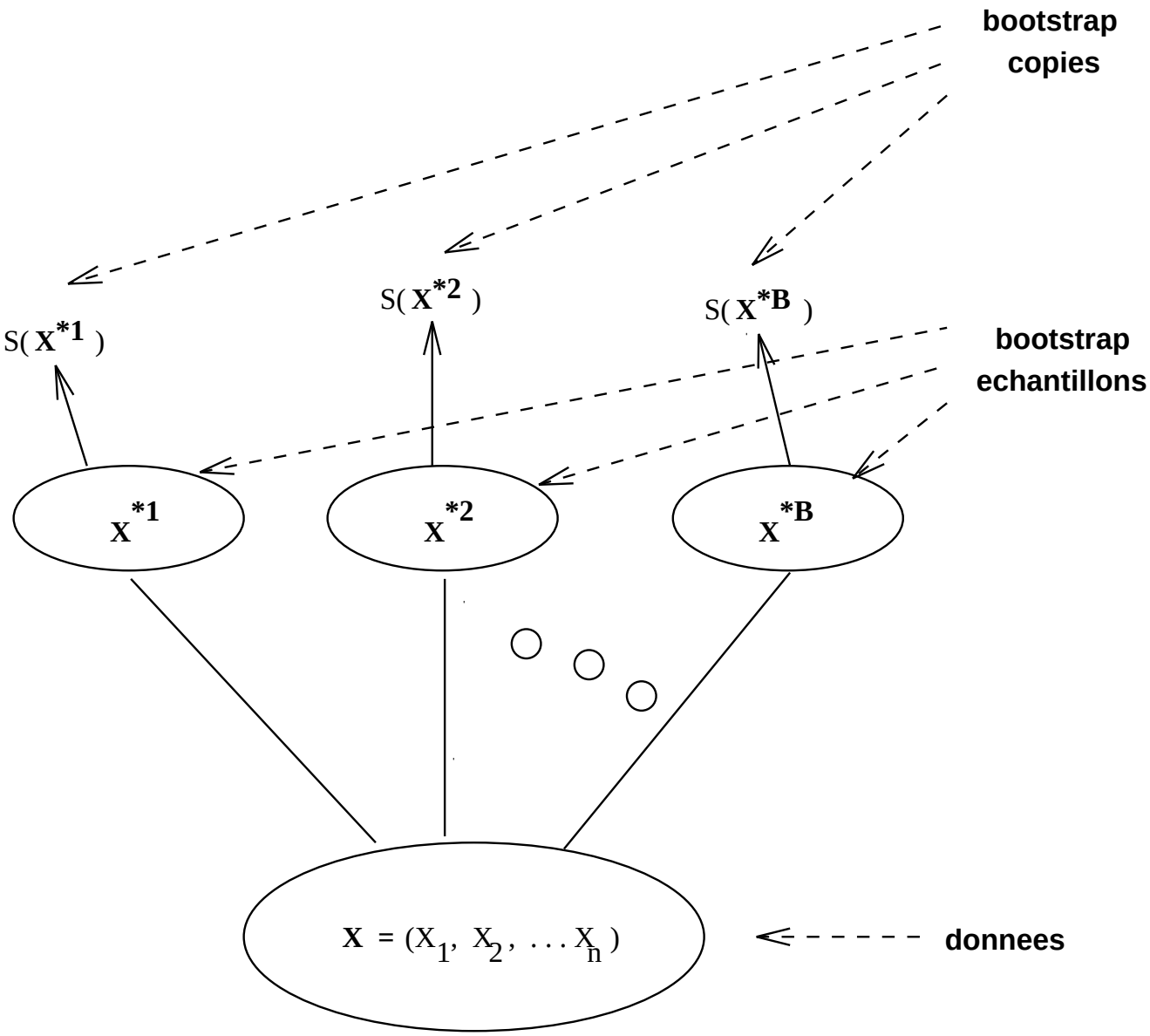
– on calcule chaque valeur $s(\mathbf{x}_b^*)$ pour chaque bootstrap.

– on calcule :

$$s(.) = \sum_{b=1}^{B} s(\mathbf{x}_b^*)/B$$

– on calcule l'estimation de l'erreur-standard :

$$\hat{\text{se}}_{\text{boot}} = \{\sum_{i=1}^{B} (s(\mathbf{x}^*) - s(.))^2/(B-1)\}^{1/2}$$


bootstrap
copies
S(X*2)
S(X*B)
S(X*1)
bootstrap
echantillons
X*1
X*2
X*B
X = (X1, X2, . . . Xn)
donnees

0.2 Quelques rappels

0.2.1 Echantillons aléatoires

ÉCHANTILLON ALÉATOIRES (rappels)

Population U : $U_1, U_2, \ldots, U_N$ de taille N.

un échantillon de taille n est un ensemble de n individus $u_1, u_2, \ldots, u_n$ sélectionnés au hasard parmi les individus de U.

en pratique on tire n entiers au hasard $j_1, j_2, \ldots, j_n$ entre 1 et N chacun ayant une probabilité $1/N$ d'être tiré. On admet les remises.

à chaque u_i on associe des mesures notées x_i, la collection

$$\mathbf{x} = (x_1, x_2, \ldots, x_n)$$

représente les *données observées*

si on avait toute la population on obtiendrait un recensement

$$X = (X_1, X_2, \ldots, X_N)$$

0.2.2 probabilités et statistiques

THÉORIE PROBABILISTE

(Population) $\rightarrow$ Propriété d'un échantillon

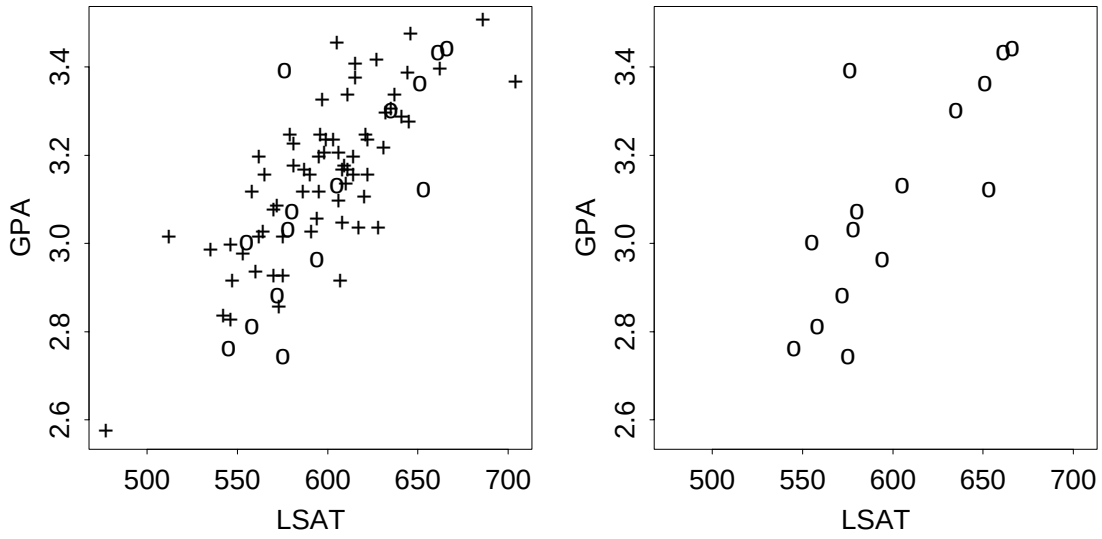
INFÉRENCE STATISTIQUE

(Observé) $\rightarrow$ Propriété de la Population

- qu'est ce qu'on apprend de X en observant $\mathbf{x}$?
- qu'elle est la précision d'une statistique calculée?

0.2.3 L'exemple des écoles

- Population : 82 écoles américaines.
- LSAT : moyenne sur le test national.
- GPA : moyenne des moyennes à la sortie.
- on tire un échantillon de 15 écoles.
- Comment à partir de l'échantillon connaître la population ?



0.3 Estimateur "Plug-in"

0.3.1 Fonction de distribution empirique

FONCTION DE DISTRIBUTION EMPIRIQUE

On a une Fonction de distribution à partir de laquelle on tire un échantillon $\mathbf{x}$

$$F \rightarrow \mathbf{x} = (x_1, x_2, \ldots, x_n)$$

La *Fonction de distribution Empirique* $\hat{F}$ donne à chaque valeur x_i la probabilité $1/n$.
En d'autres termes :

$$\widehat{Prob}\{A\} = Prob_{\widehat{F}}\{A\} = \#\{x_i \in A\}/n$$

Exemple :

$$A = \{(y, z) : 0 < y < 600, 0 < z < 3.00\}$$

$$Prob(A) = 16/82 = 0.195$$

$$Prob_{\widehat{F}}\{A\} = 5/15 = 0.333$$

0.3.2 Comment sont liés y et z ?

Comment sont liés y et z ?

Si on regarde toute la population : on connaît exactement les espérances des deux variables et on peut calculer **LA** statistique

$$\mathrm{corr}(y,z) = \frac{\Sigma_{i=1}^{82}(Y_j-\mu_y)(Z_j-\mu_z)}{[\Sigma_{i=1}^{82}(Y_j-\mu_y)^2\,\Sigma_{i=1}^{82}(Z_j-\mu_z)^2]^{1/2}}$$

$$\mu_y = 597.5 \text{ et } \mu_z = 3.13$$

$$\mathrm{corr}(y,z) = 0.761$$

Ce n'est que de l'arithmétique !

Mais le coefficient de corrélation de l'échantillon :

$$\widehat{\mathrm{corr}}(y,z) = \frac{\Sigma_{i=1}^{15}(y_j-\hat{\mu}_y)(z_j-\hat{\mu}_z)}{[\Sigma_{i=1}^{15}(y_j-\hat{\mu}_y)^2\,\Sigma_{i=1}^{15}(z_j-\hat{\mu}_z)^2]^{1/2}}$$

$$\hat{\mu}_y = 600.3 \text{ et } \hat{\mu}_z = 3.09$$

$$\widehat{\mathrm{corr}}(y,z) = 0.776$$

Ce n'est qu'une estimation de $\mathrm{corr}(y,z)$!

0.3.3 Le principe "je branche"

LE PRINCIPE "JE BRANCHE"

- *Le paramêtre :* $\theta = t(F)$
  C'est une fonction de la distribution de probabilité.
- *L'estimateur "plug-in" :* $\hat{\theta} = t(\hat{F})$
  C'est la même fonction utilisée pour $\hat{F}$ que pour F.

$$\begin{aligned} \hat{\theta} &= \text{résumé statistique} \\ &= \text{la statistique} \\ &= \text{l'estimateur} \\ &= \textbf{LA} \text{ sortie d'un listing} \end{aligned}$$

Oui , mais comment $\hat{\theta}$ approche-t-il θ?

- Quel est le biais?
- Quelle est l'erreur "standard" ?

C'est tout ce que tente de résoudre LE CYRANO ou BOOTSTRAP

0.4 Erreurs standards

0.4.1 Erreur-standard d'une moyenne

ERREUR-STANDARD D'UNE MOYENNE

Soit x une variable aléatoire de distribution F :
Espérance et variance de F :

$$\mu_F = E_F(x),\ \sigma_F^2 = \mathrm{var}_F(x) = E_F[(x-\mu_F)^2]$$

On note :

$$x \sim (\mu_F, \sigma_F^2)$$

On tire un échantillon $\mathbf{x}$ de taille n à partir de F,

$$F \rightarrow \mathbf{x} = (x_1, x_2, \ldots, x_n)$$

La moyenne $\overline{x} = \Sigma_{i=1}^n x_i/n$ de l'échantillon a pour espérance μ_F et pour variance σ_F^2/n

$$\overline{x} \sim (\mu_F, \sigma_F^2/n)$$

l'erreur-standard de la moyenne $\overline{x}$:

$$\mathrm{se}_F(\overline{x}) = \mathrm{se}(\overline{x}) = \sqrt{\mathrm{var}_F(x)} = \sigma_F/n$$

Qu'est ce que ca veut dire ?

0.4.2 Théorème central limite

THÉORÈME CENTRAL LIMITE

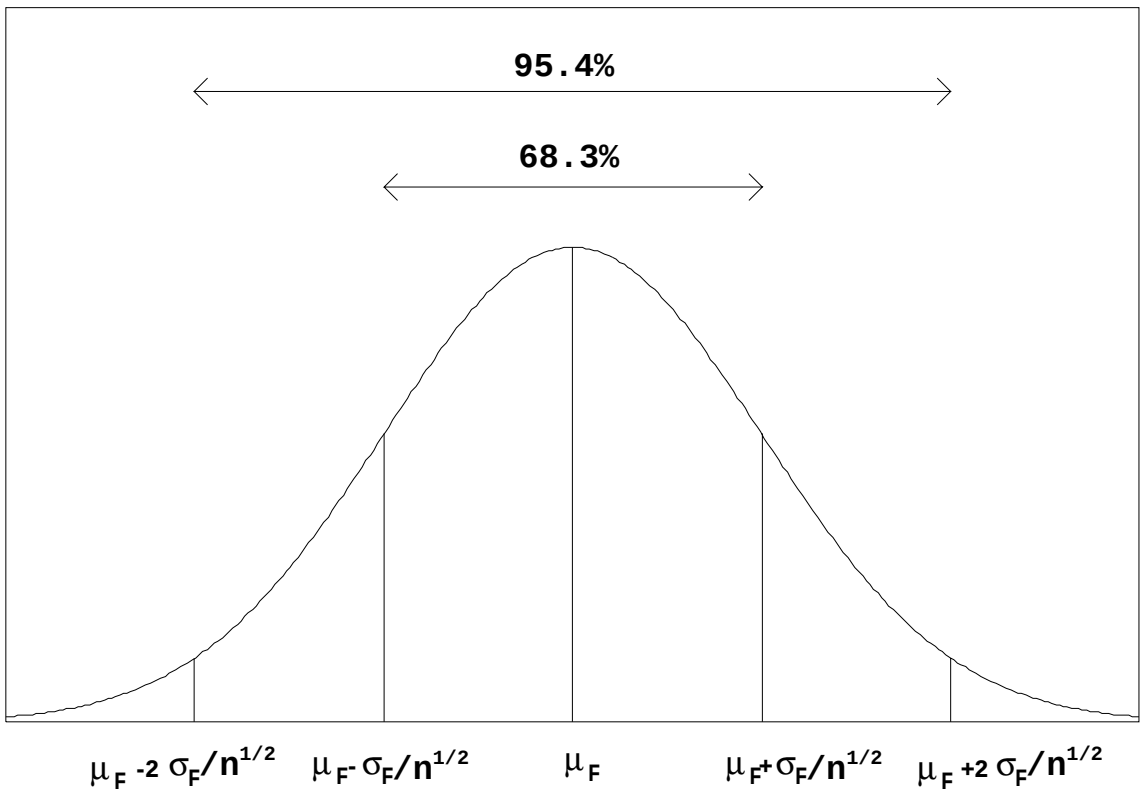
Quand n est "assez grand", $\overline{x}$ suit une loi Normale d'espérance μ_F et de variance σ^2/n.

$$\overline{x} \asymp \mathrm{N}(\mu_F, \sigma^2, n)$$

En d'autres termes :

$$\mathrm{Prob}\{|\overline{x} - \mu_F| < \frac{\sigma}{\sqrt{n}}\} \doteq 0.683$$

$$\mathrm{Prob}\{|\overline{x} - \mu_F| < 2\frac{\sigma}{\sqrt{n}}\} \doteq 0.954$$



0.4.3 Tirages à pile ou face

TIRAGES A PILE OU FACE

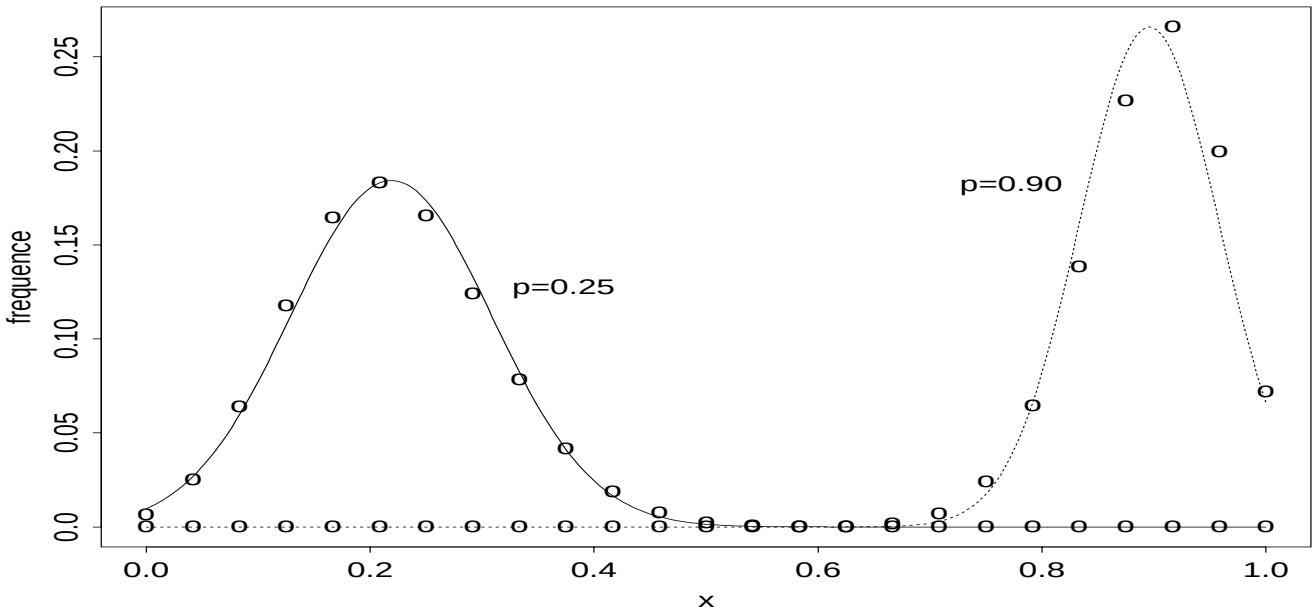
$$\text{Prob}_F\{x = 1\} = p \text{ et } \text{Prob}_F\{x = 0\} = 1 - p$$

On tire n fois la pièce, soit s le nombre de fois où on tire *"pile"*

$s = \Sigma_{i=1}^{n} x_i$ suit une binomiale.

la moyenne $\overline{x} = s/n = \hat{p}$ est l'estimateur "plug-in" de p.

$$\hat{p} =\sim (p, p(1-p)/n)$$



Et encore on ne connaît pas σ, on a juste une estimation

0.5 Estimateurs bootstrap de l'erreur standard

0.5.1 Estimateur bootstrap non-papramètrique de l'erreur standard

ESTIMATEUR BOOTSTRAP NON-PARAMÈTRIQUE DE L'ERREUR STANDARD

On a tiré un échantillon $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ d'une fonction de distribution inconnue F.

$$F \rightarrow \mathbf{x} = (x_1, x_2, \ldots, x_n)$$

On veut estimer un paramêtre $\theta = t(F)$ à partir de $\mathbf{x}$

On calcule un estimateur $\hat{\theta} = s(\mathbf{x})$

$\hat{F}$ est la fonction de distribution qui donne la probabilité $1/n$ à chaque x_i

on tire des *échantillons Bootstrap* à partir de $\hat{F}$

$$\hat{F} \rightarrow \mathbf{x}_b^* = (x_1^*, x_2^*, \ldots, x_n^*)$$

on calcule la copie bootstrap $\hat{\theta}^* = s(\mathbf{x}^*)$

l'estimation bootstrap idéale de l'erreur-standard $\mathrm{se}_F(\hat{\theta})$ est l'erreur-standard de $\hat{\theta}$ pour des ensembles de données de taille n tirés suivant $\hat{F}$, c'est-à-dire $\mathrm{se}_{\hat{F}}(\hat{\theta}^*)$

0.5.2 Algorithme d'estimation des erreurs-standards

Algorithme d'estimation des erreurs-standards

1. Tirer B échantillons bootstrap $\mathbf{x}^{*1}, \mathbf{x}^{*2}, \ldots, \mathbf{x}^{*B}$ à partir de $\mathbf{x}$
2. calculer la copie bootstrap $\hat{\theta}^*(b) = s(\mathbf{x}^{*b})$
3. calculer l'erreur-standard pour les B copies

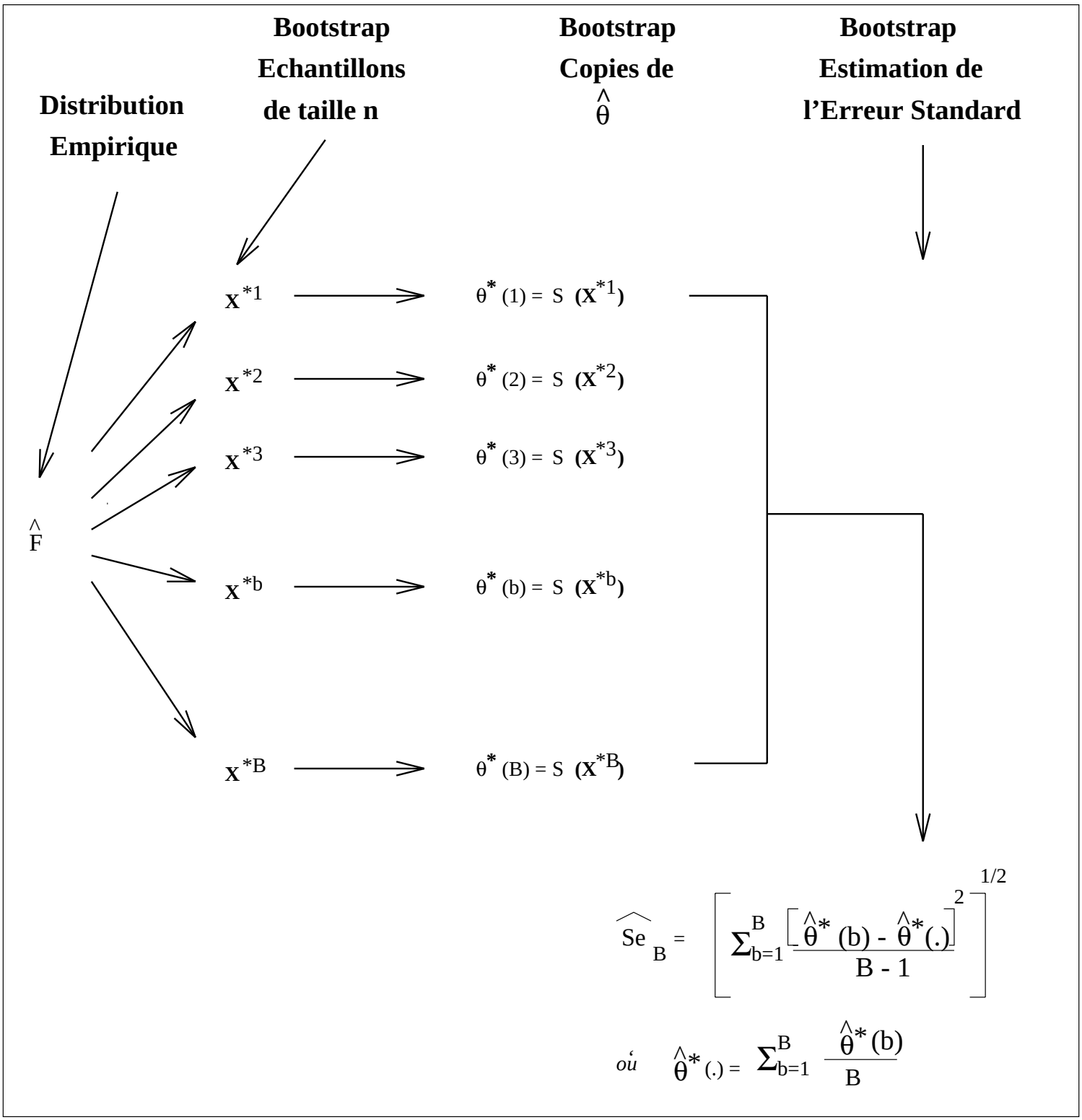
$$\hat{\text{se}}_B = \{\sum_{b=1}^{B} (\hat{\theta}^*(b) - \hat{\theta}^*(.))^2/(B-1)\}^{1/2}$$

avec $\hat{\theta}^*(.) = \Sigma_{b=1}^{B} \hat{\theta}^*(b)/B$.

l'estimation bootstrap non-paramétrique de l'erreur-standard $\hat{\text{se}}_B$ a pour limite $\text{se}_F(\hat{\theta})$ quand B tend vers l'infini.

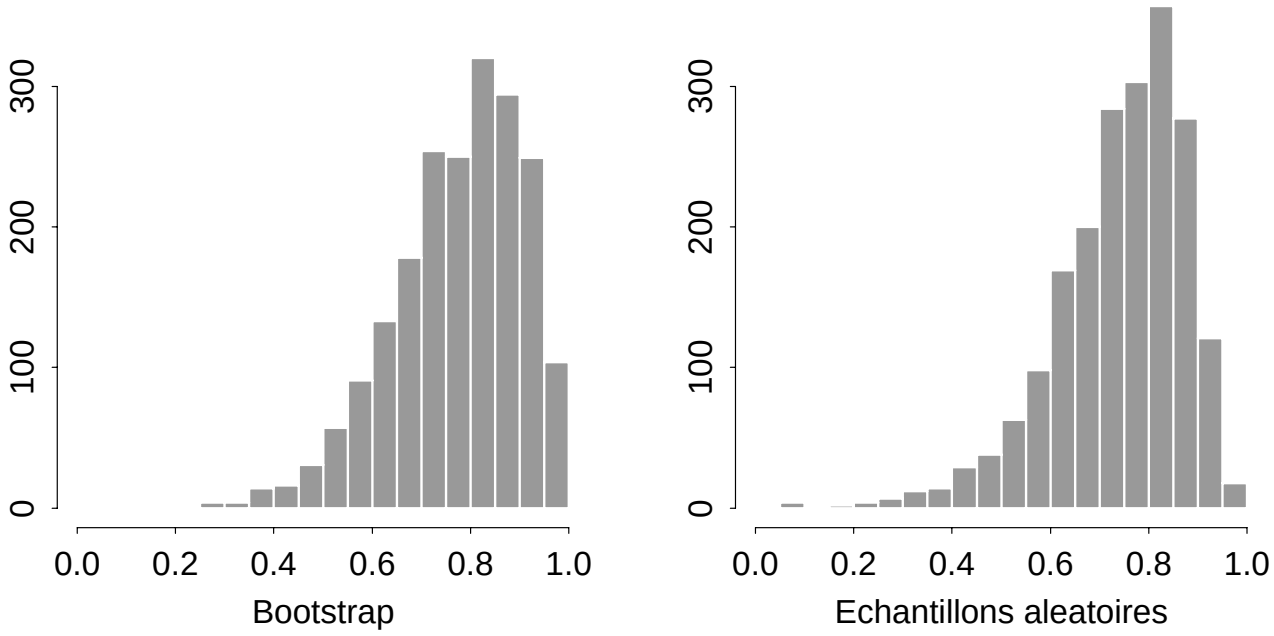
0.5.3 Algorithme pour l'estimateur bootstrap de l'erreur-standard

ALGORITHME POUR CALCULER L'ESTIMATEUR BOOTSTRAP DE L'ERREUR STANDARD



0.5.4 un exemple

UN EXEMPLE :
Le coefficient de corrélation pour les écoles



B :	25	50	100	200	400	800	1600	3200
$\hat{\text{se}}_B$:	0.140	0.142	0.151	0.143	0.141	0.137	0.133	0.132

0.5.5 Combien de bootstraps?

COMBIEN DE BOOTSTRAP?

Ca dépend!

Règles "pifométriques"

- B=25 à B=50 : pour obtenir un début d'information.
- B=200 : pour estimer l'erreur-standard.
- B=500 : pour l'évaluation d'intervalles de confiance.

0.5.6 Estimateur bootstrap paramétrique de l'erreur-standard

ESTIMATEUR BOOTSTRAP PARAMÉTRIQUE DE L'ERREUR STANDARD

On suppose qu'on a une estimation $\hat{F}_{\text{par}}$ d'un modèle paramétrique de F.

l'estimateur Bootstrap paramétrique de l'erreur-standard est $\text{se}_{\hat{F}_{\text{par}}}(\hat{\theta}^*)$

Exemple : les écoles.
On suppose que les 2 variables suivent une loi binormale de moyenne $(\overline{y}, \overline{z})$ et de matrice de covariance :

$$\frac{1}{14}\begin{pmatrix} \Sigma(y_i-\overline{y})^2 & \Sigma(y_i-\overline{y})(z_i-\overline{z}) \\ \Sigma(y_i-\overline{y})(z_i-\overline{z}) & \Sigma(z_i-\overline{z})^2 \end{pmatrix}$$

On tire B échantillons de taille 15 qui suivent cette loi et on calcule les copies bootstrap, puis l'estimateur bootstrap de l'erreur-standard.

0.5.7 Avant le bootstrap

AVANT LE BOOTSTRAP

- QUE DES MATHS !
- Quelques distributions
- Quelques statistiques
- Exemple : la corrélation. Si F est un loi bi-normale alors

$$\hat{\text{se}}_{\text{normal}} = (1 - \hat{\text{corr}}^2)/\sqrt{n-3}$$

AVEC LE BOOTSTRAP

- QUE DES CALCULS INFORMATIQUES !
- On est libre des hypothèses
- On est plus près des données

0.6 Exemple d'utilisation

0.6.1 L'analyse factorielle

EXEMPLE D'ANALYSE FACTORIELLE

88 étudiants ont passé 5 examens en mathématiques.

Matrice de données : 88 lignes $\mathbf{x}_i = (x_{i1}, x_{i2}, x_{i3}, x_{i4}, x_{i5})$ et 5 colonnes.

moyenne empirique $\overline{x} = (38.95, 50.59, 50.60, 46.68, 42.31)$

matrice empirique de covariance : G

Calcul des 5 valeurs propres de G :

$$\begin{aligned} \hat{\lambda}_1 &= 679.2 \\ \hat{\lambda}_2 &= 199.8 \\ \hat{\lambda}_3 &= 102.6 \\ \hat{\lambda}_4 &= 83.7 \\ \hat{\lambda}_5 &= 31.8 \end{aligned}$$

0.6.2 Pourquoi une acp ?

POURQUOI UNE ACP ?

Si il existe une seule valeur Q_i pour chaque étudiant qui pourrait le "résumer" et 5 valeurs $\mathbf{v} = (v_1, v_2, v_3, v_4, v_5)$ avec

$$\mathbf{x}_i = Q_i \mathbf{v}$$

alors seul $\hat{\lambda}_1$ est positif, les autres sont nuls.

$$\hat{\theta} = \hat{\lambda}_1 / \sum_{i=1}^{5} \hat{\lambda}_i$$

$\hat{\theta}$ est-il égal à 1 ?

$$\hat{\theta} = 0.619$$

Quelle est la précision de $\hat{\theta}$

0.6.3 Le bootstrap

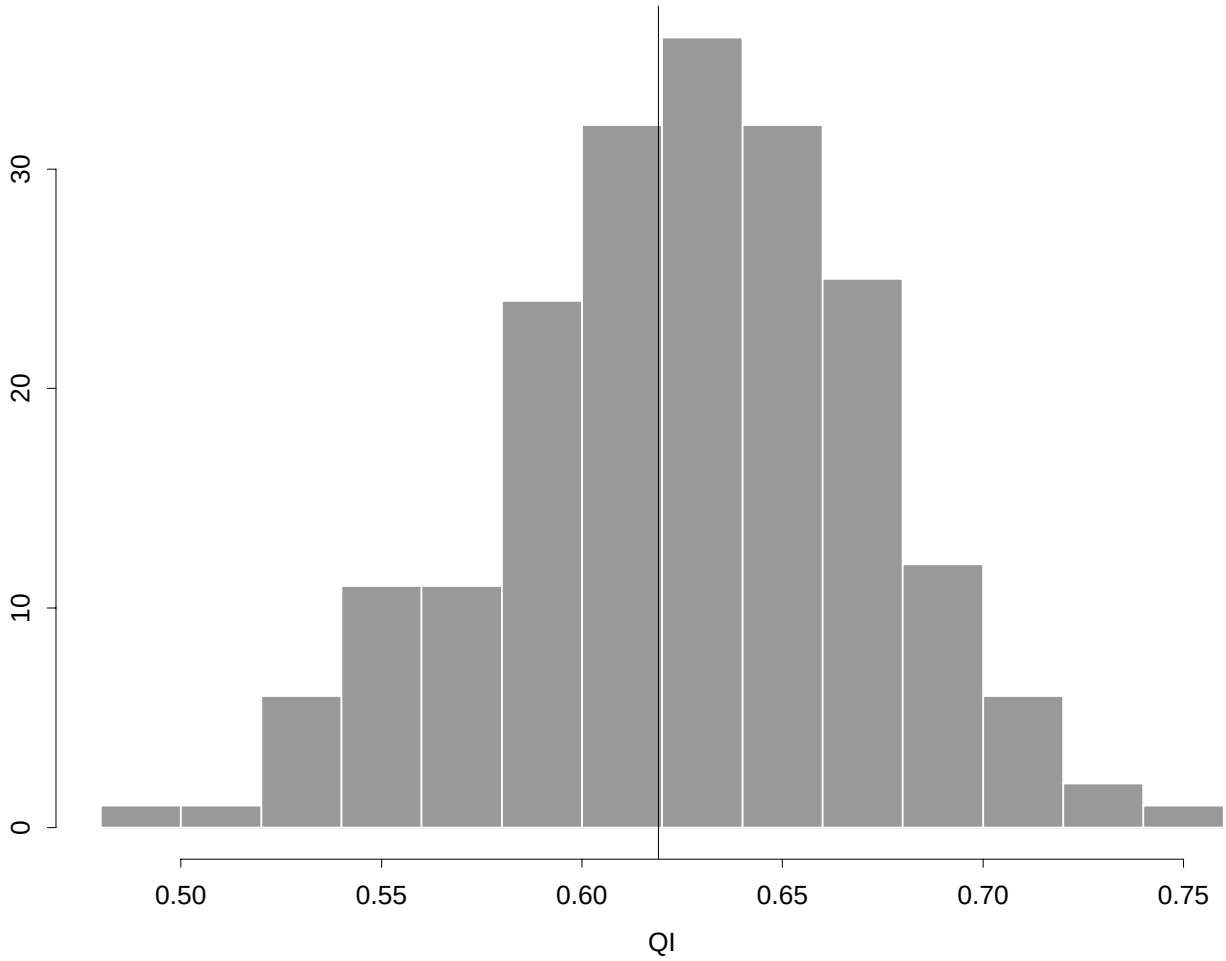
LE BOOTSTRAP

Chaque échantillon bootstrap sera une matrice $\mathbf{X}^*$ de 5 colonnes et 88 lignes tirée au hasard parmi les $\mathbf{x}_i$.

On calcule $\mathbf{G}^*$

On calcule les 5 valeurs propres : $\hat{\lambda}_i$

On calcule : $\hat{\theta}^* = \hat{\lambda}_1^* / \Sigma_{i=1}^5 \hat{\lambda}_i^*$


0
10
20
30
0.50
0.55
0.60
0.65
0.70
0.75
QI

$$\overline{\theta}^* = 0.625$$

$$\hat{\text{se}}_{200} = 0.047$$

L'intervalle de confiance standard pour la vraie valeur de θ avec une probabilité $(1 - 2\alpha)$ est :

$$\theta \in \hat{\theta} \pm z^{(1-\alpha)}.\hat{\text{se}}$$

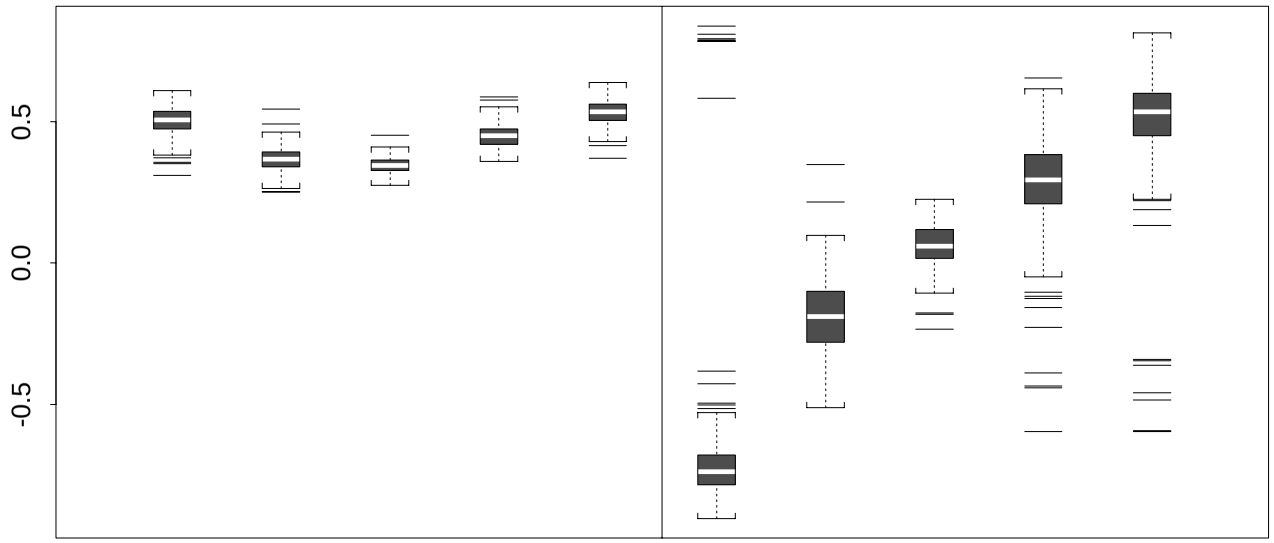
où $z^{(1-\alpha)}$ est le $100(1-\alpha)$-ième percentile de la loi normale centrée réduite.
Ici :

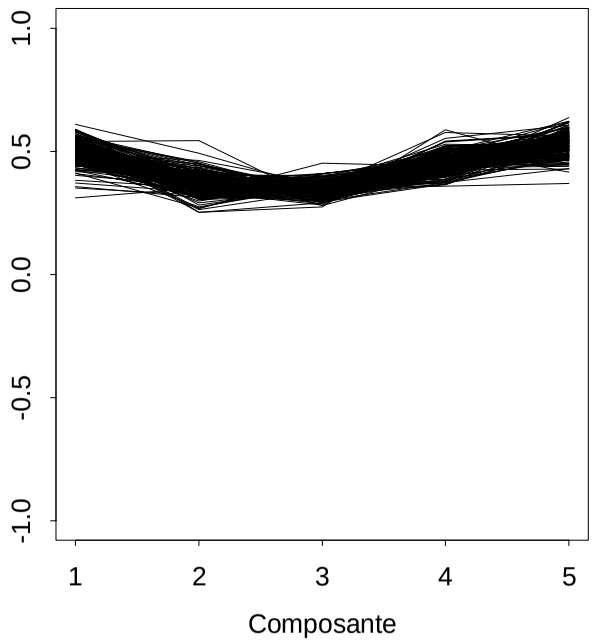
$$\theta \in 0.619 \pm 0.047 = [0.572, 0.666]$$
avec une probabilité 0.683

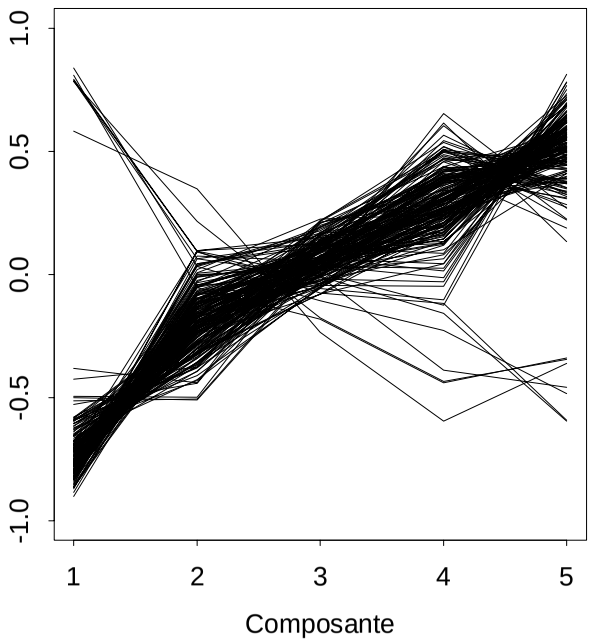
$$\theta \in 0.619 \pm 1.645 * 0.047 = [0.542, 0.696]$$
avec une probabilité 0.900

0.6.4 Et les vecteurs propres

POURQUOI PAS LA MÊME CHOSE AVEC LES VECTEURS PROPRES?



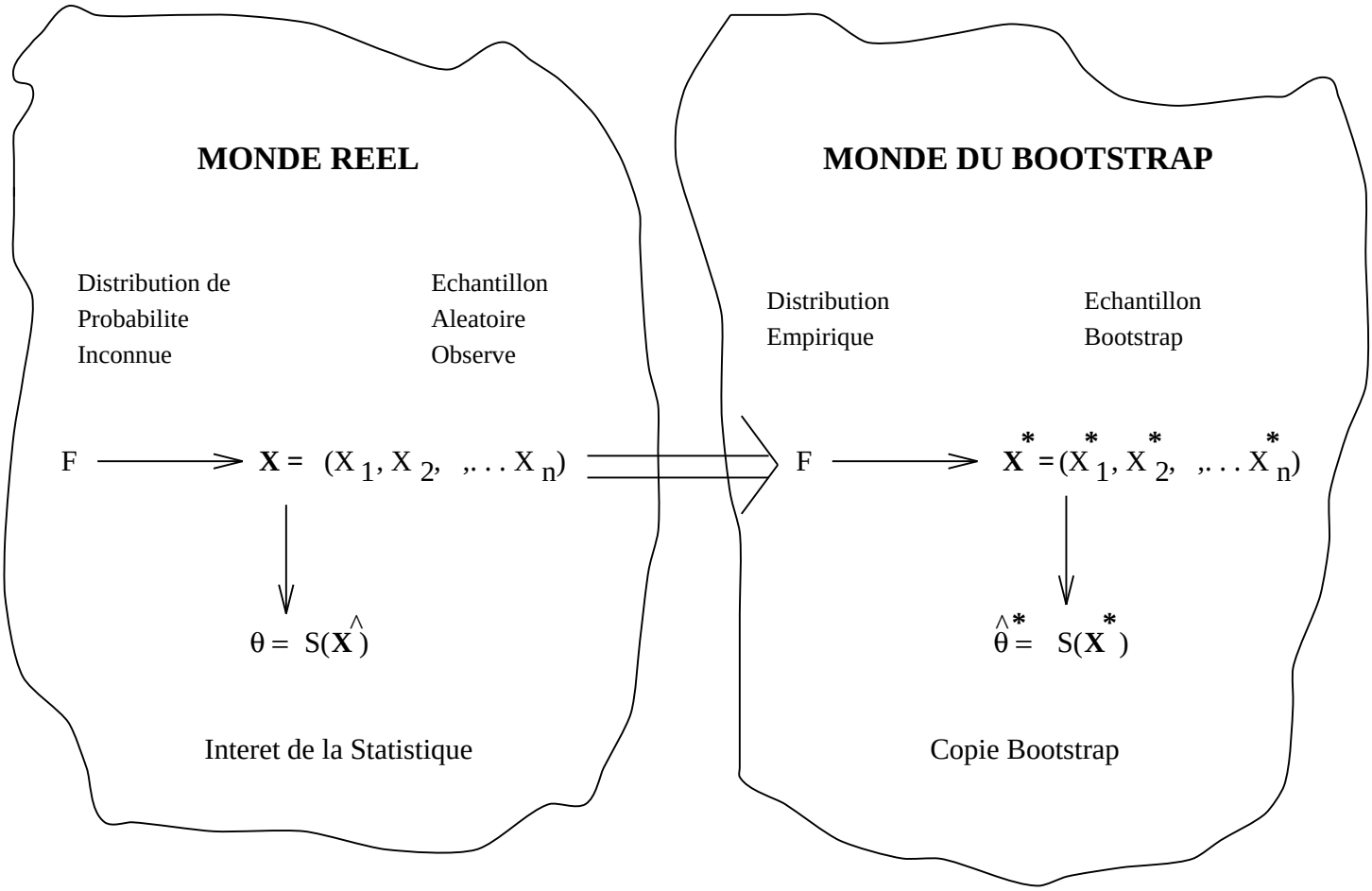

1.0
0.5
0.0
-0.5
-1.0
1
2
3
4
5
Composante


1.0
0.5
0.0
-0.5
-1.0
1
2
3
4
5
Composante

0.7 Structures de données plus complexes

0.7.1 Généralités

DES STRUCTURES DE DONNÉES PLUS COMPLEXES



- Le Point Crucial : $\Longrightarrow$
- Comment calculer une estimation $\hat{F}$ de F à partir des données $\mathbf{x}$?
- On note :

$$P \longrightarrow \mathbf{x}$$

"Du modèle probabiliste P on a tiré $\mathbf{x}$"

0.7.2 Deux échantillons

DEUX ÉCHANTILLONS

Les souris

Soit F la distribution des valeurs pour le groupe traité
Soit G la distribution des valeurs pour le groupe de contrôle

$$P = (F, G)$$

$\mathbf{z} = (z_1, z_2, \ldots, z_m)$ les valeurs du groupe traité
$\mathbf{y} = (y_1, y_2, \ldots, y_n)$ les valeurs du groupe de contrôle

$$\mathbf{x} = (\mathbf{z}, \mathbf{y})$$

$$P \longrightarrow \mathbf{x}$$

c'est

$$F \longrightarrow \mathbf{z} \text{ indépendamment de } G \longrightarrow \mathbf{y}$$

0.7.3 Bootstrap

BOOTSTRAP

$\hat{F}$ et $\hat{G}$ les distributions empiriques basées sur $\mathbf{z}$ et $\mathbf{y}$

$$\hat{P} = (\hat{F}, \hat{G})$$

$$\hat{P} \longrightarrow \mathbf{x}^*$$

c'est

$$\hat{F} \longrightarrow \mathbf{z}^* \text{ indépendamment de } \hat{G} \longrightarrow \mathbf{y}^*$$

– échantillons bootstrap

$$\mathbf{x}^* = (\mathbf{z}^*, \mathbf{y}^*) = (z_{i_1}, z_{i_2}, \ldots, z_{i_7}, z_{j_1}, z_{j_2}, \ldots, z_{j_9})$$

où $(i_1, i_2, \ldots, i_7)$ est un échantillon de taille 7 tiré parmi les entiers 1,2,. . . ,7 et $(j_1, j_2, \ldots, j_9)$ est un échantillon de taille 9 tiré parmi les entiers 1,2,. . . ,9 indépendamment.

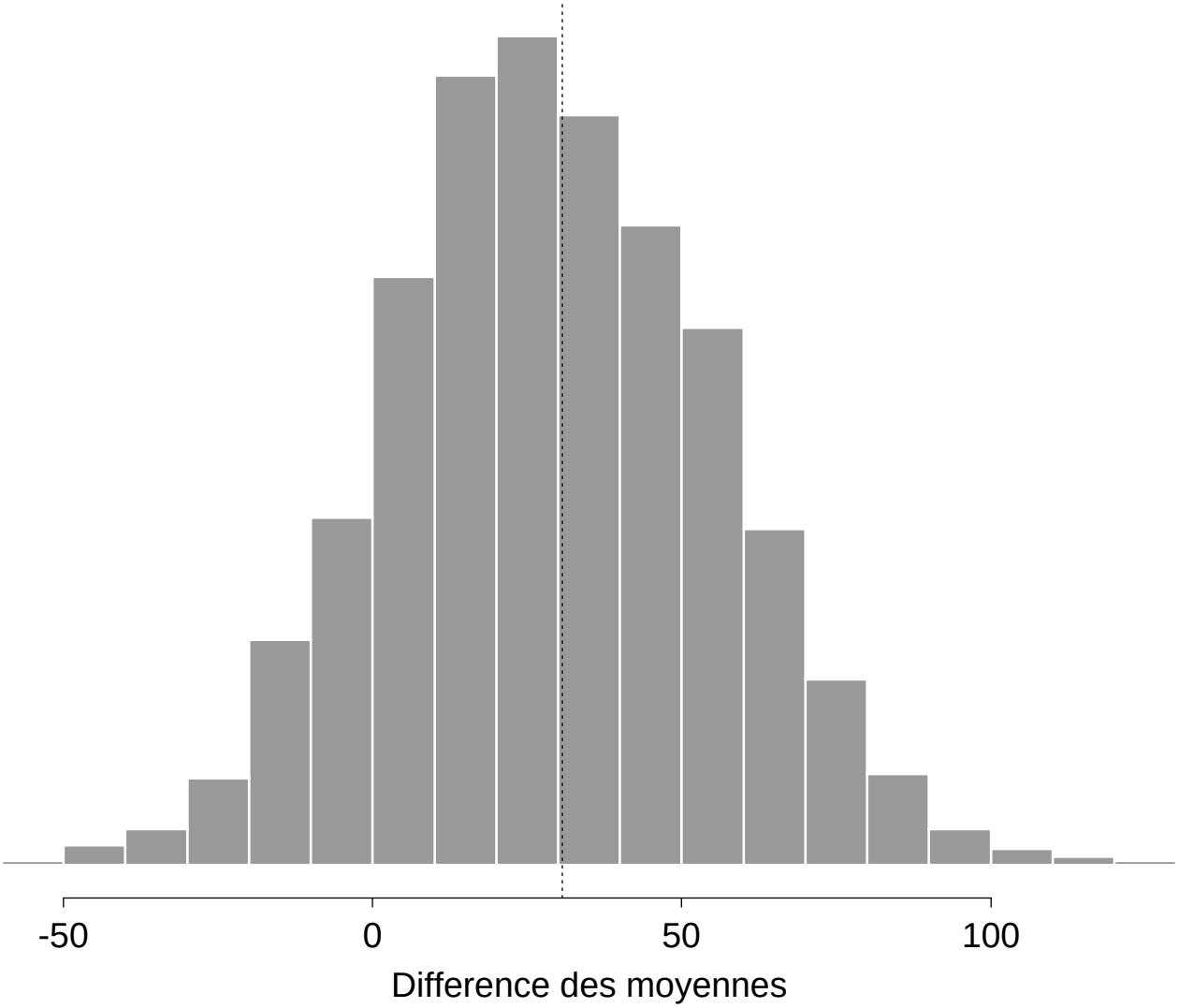
– paramêtre : $\theta = \mu_z - \mu_y = E_F(z) - E_G(y)$

– statistique : $\hat{\theta} = \hat{\mu}_z - \hat{\mu}_y = \overline{z} - \overline{y} = 30.63$

– copie bootstrap $\theta^* = \overline{z}^* - \overline{y}^*$

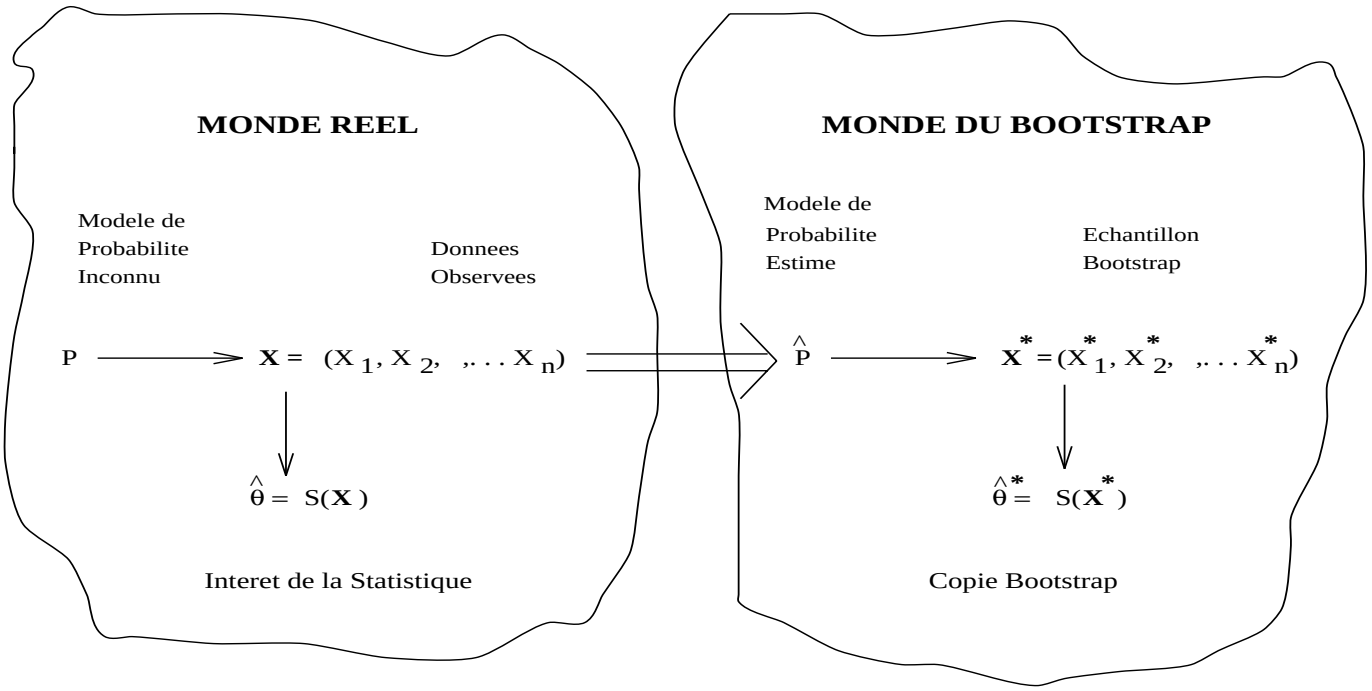
– estimation bootstrap de l'erreur-standard : $\hat{\text{se}}_{1400} = 26.85$

– rapport : $\hat{\theta}/\hat{\text{se}}_{1400} = 1.14$ trop petit pour conclure à un effet !


-50
0
50
100
Difference des moyennes

0.7.4 Structures de données plus générales

STRUCTURES DE DONNÉES PLUS GÉNÉRALES



1. On doit estimer le modèle probabiliste P à partir des données $\mathbf{x}$:

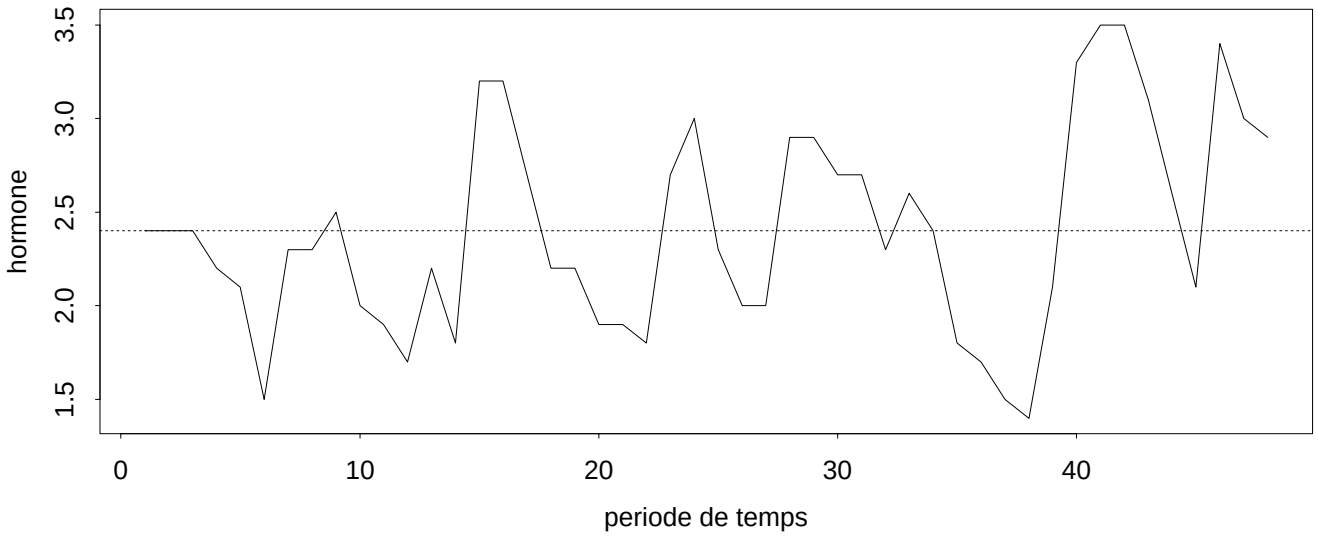
$$\mathbf{x} \Longrightarrow \hat{P}$$

2. On doit simuler des données bootstrap $\mathbf{x}^*$ à partir de $\hat{P}$:

$$\hat{P} \longrightarrow \mathbf{x}^*$$

0.7.5 Exemple : série chronologique

EXEMPLE : SÉRIE CHRONOLOGIQUE



– A chaque temps t on mesure une hormone y_t

– modèle le plus simpliste : auto-régressif d'ordre 1.

$$z_t = y_t - E_F(y) = y_t - \mu$$

$$AR(1) : z_t = \beta z_{t-1} + \epsilon_t pour 1 < t \leq 48 \text{ avec } -1 < \beta < 1$$

– les ϵ_t sont un échantillon tiré d'une loi de distribution F

$$F \longrightarrow (\epsilon_2, \epsilon_3, \ldots, \epsilon_{48})$$

avec $E_F(\epsilon) = 0$.

0.7.6 Comment utiliser le bootstrap ?

COMMENT UTILISER LE BOOTSTRAP ?

- Estimation du paramêtre β.
  - exemple : par les moindres carrés.

$$\hat{\beta} / \min_b \sum_2^{48} (z_t - b z_{t-1})^2$$

$$\hat{\beta} = 0.586$$

- Précision de l'estimateur $\hat{\beta}$???
  - bootstrap
- Quel est le modèle probabiliste P ??
  - $P = (\beta, F)$
- Estimation de : $\mathbf{x} \Longrightarrow \hat{P}$
  - on estime β par $\hat{\theta}$.
  - soit $\hat{\epsilon}_t = z_t - \hat{\beta} z_{t-1}$ $\hat{F}$ définit une probabilité 1/47 sur chaque $\hat{\epsilon}_t = z_t$

– Tirage du bootstrap : $\hat{P} \longrightarrow \mathbf{x}^*$

– On pose $z_1 = y_1 - \overline{y}$ C'est une constante comme n=48 !

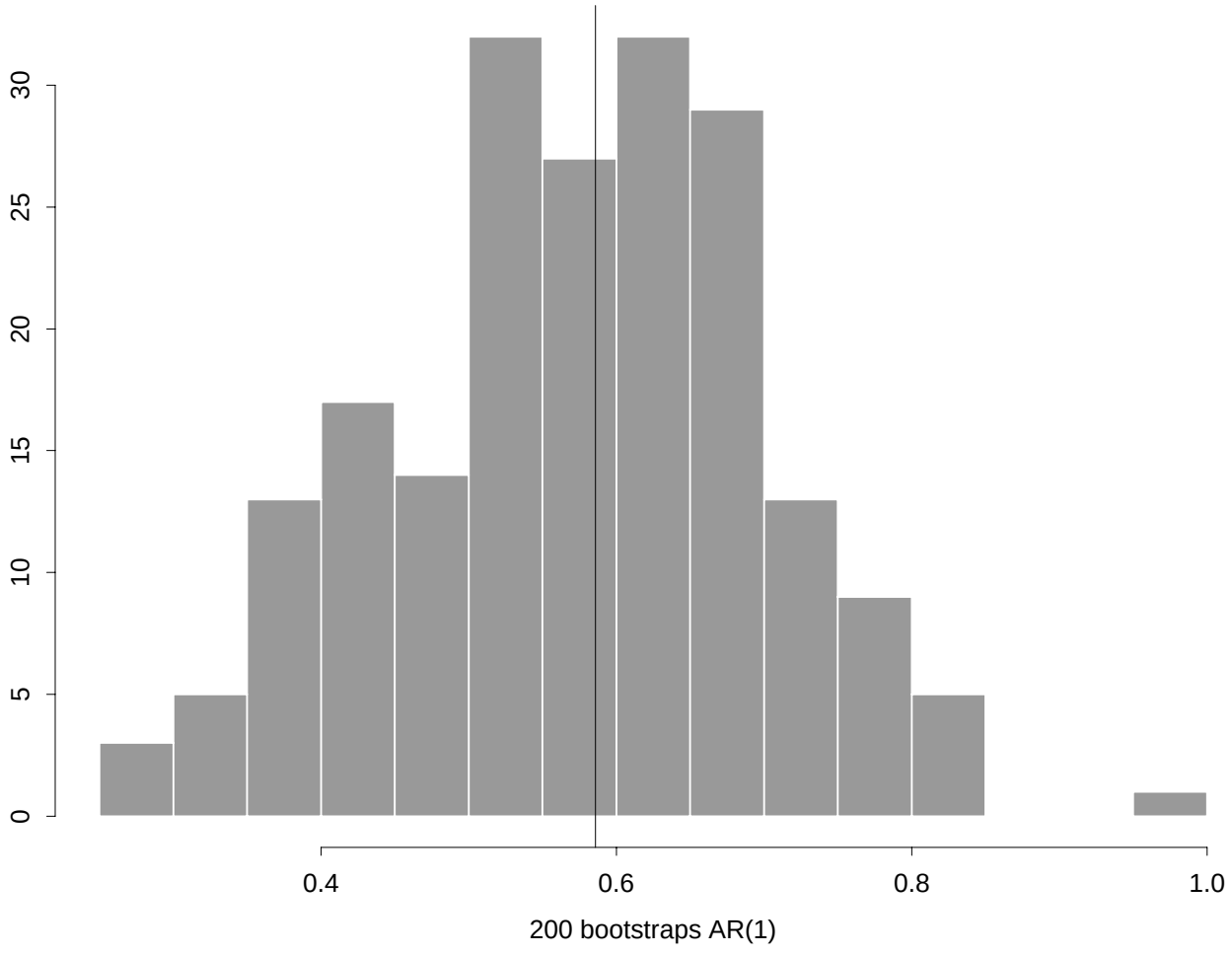
– On tire les ϵ_t^* à partir de $\hat{F}$:

$$\hat{F} \longrightarrow (\epsilon_2^*, \epsilon_3^*, \ldots, \epsilon_{48}^*)$$

– On pose :

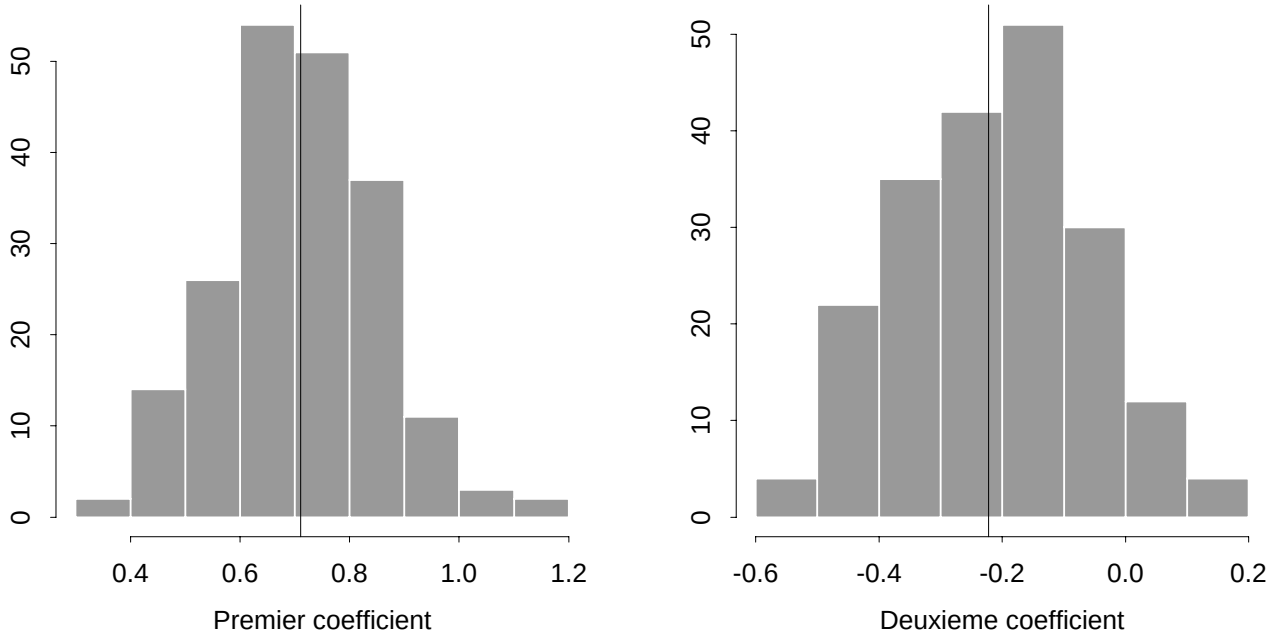
$$\begin{aligned} z_2^* &= \hat{\beta} z_1 + \epsilon_2^* \\ z_3^* &= \hat{\beta} z_2 + \epsilon_3^* \\ &\vdots \quad\quad \vdots \\ z_{48}^* &= \hat{\beta} z_{47} + \epsilon_{48}^* \end{aligned}$$

– Pour chaque bootstrap on calcule $\hat{\beta}^*$



Est ce vraiment un AR(1)?

On teste avec un AR(2)



Autre méthode pour les séries chronologiques : bootstrap par blocs.

0.8 Régression linéaire

0.8.1 Présentation du problème

RÉGRESSION LINÉAIRE

$$\mathbf{x}_i = (\mathbf{c}_i, y_i)$$

$$\mathbf{c}_i = (c_{i1}, c_{i2}, \ldots, c_ip)$$

$$y_i : \text{ réponse } \mathbf{c}_i : \text{ prédicteurs}$$

$$\mu_i = \mathrm{E}(y_i|\mathbf{c}_i) = \mathbf{c}_i\beta = \sum_{i=1}^{p} c_{ij}\beta_j$$

Estimateur des moindres carrés :

$$\hat{\beta} / \min_b \sum_{i=1}^{n} (y_i - \mathbf{c}_i\mathbf{b})^2$$

$$\mathbf{C} = \begin{pmatrix} \mathbf{c}_1 \\ \mathbf{c}_2 \\ \vdots \\ \mathbf{c}_n \end{pmatrix}$$

$$\hat{\beta} = (\mathbf{C}^T\mathbf{C})^{-1}\mathbf{C}^T\mathbf{y}$$

0.8.2 modèle probabiliste

modèle probabiliste

Modèle : $y_i = \mathbf{c}_i\beta + \epsilon_i$ ϵ_i suit une loi de distribution F.

$$\sigma_F^2 = \mathrm{var}_F(\epsilon) \text{ et } \mathbf{G} = \mathbf{C}^T\mathbf{C}$$

alors :

$$\mathrm{se}(\hat{\beta}_j) = \sigma_F\sqrt{\mathbf{G}_{jj}} \text{ où } \mathbf{G}_{jj} \text{ élément diagonal de } \mathbf{G}^{-1}$$

Mais on ne connaît pas σ_F : estimation

1. $\hat{\sigma}_F = \Big[\mathrm{RSE}(\hat{\beta})/n\Big]^{1/2}$
2. $\overline{\sigma}_F = \Big[\mathrm{RSE}(\hat{\beta})/(n-p)\Big]^{1/2}$ estimation non biaisée.

$$\hat{\mathrm{se}}(\hat{\beta}_j) = \hat{\sigma}_F\sqrt{\mathbf{G}_{jj}} \text{ où } \overline{\mathrm{se}}(\hat{\beta}_j) = \overline{\sigma}_F\sqrt{\mathbf{G}_{jj}}$$

0.8.3 Bootstrap

bootstrap

1. mécanisme probabiliste: $P \longrightarrow \mathbf{x}$.

$$2 \text{ composantes}: P = (\beta, F)$$

2. estimation de $\hat{P}$ $\mathbf{x} \Longrightarrow \hat{P}$

   – estimation de $\hat{\beta}$
   – Soit $\hat{\epsilon}_i = y_i - \mathbf{c}_i\hat{\beta}$

$$\hat{F}: \text{probabilité } 1/n \text{ sur chaque } \hat{\epsilon}_i$$
$$\hat{P} = (\hat{\beta}, \hat{F})$$

3. bootstrap : $\hat{P} \longrightarrow \mathbf{x}^*$

   – Génération : $\hat{F} \longrightarrow (\epsilon_2^*, \epsilon_3^*, \ldots, \epsilon_n^*)$
   – Puis :

$$y_i^* = \mathbf{c}_i\hat{\beta} + \epsilon_i^*\mathbf{x}_i^* = (\mathbf{x}_i, y_i^*)$$

4. estimation bootstrap :

$$\hat{\beta}^* = (\mathbf{C}^T\mathbf{C})^{-1}\mathbf{C}^T\mathbf{y}^*$$

5. estimation bootstrap de l'erreur-standard :

$$\begin{aligned} \text{var}(\hat{\beta}^*) &= (\mathbf{C}^T\mathbf{C})^{-1}\mathbf{C}^T\text{var}(\mathbf{y}^*)\mathbf{C}(\mathbf{C}^T\mathbf{C})^{-1} \\ &= \hat{\sigma}_F^2(\mathbf{C}^T\mathbf{C})^{-1} \end{aligned}$$

car $\text{var}(\mathbf{y}^*) = \hat{\sigma}_F^2 \text{I}$. Ainsi :

$$\hat{\text{se}}_\infty(\hat{\beta}_j) = \hat{\sigma}_F \sqrt{\mathbf{G}_{jj}} = \hat{\text{se}}(\hat{\beta}_j)$$

c'est le même!

0.8.4 bootstrap par paires ou par résidus ?

bootstrap par paires ou par résidus ?

1. par paires :

$$\mathbf{x}_i = \{(\mathbf{c}_{i_1}, y_{i_1}), (\mathbf{c}_{i_2}, y_{i_2}), \ldots, (\mathbf{c}_{i_n}, y_{i_n})\}$$

2. bootstrap des résidus :

$$\mathbf{x}_i = \{(\mathbf{c}_1, \mathbf{c}_1\hat{\beta} + \hat{\epsilon}_{i_1}), (\mathbf{c}_2, \mathbf{c}_2\hat{\beta} + \hat{\epsilon}_{i_2}), \ldots, (\mathbf{c}_n, \mathbf{c}_n\hat{\beta} + \hat{\epsilon}_{i_n})\}$$

Quel est le meilleur ?

Ça dépend de la confiance qu'on a dans le modèle linéaire

– le modèle suppose que ϵ_i ne dépend pas de $\mathbf{c}_i$: c'est-à-dire que F est le même pour tout $\mathbf{c}_i$.

– "Par paires" est moins sensible aux hypothèses.

0.9 Estimation du biais

0.9.1 Présentation du problème

ESTIMATION DU BIAIS

- Une fonction de distribution inconnue : F

$$F \rightarrow \mathbf{x} = (x_1, x_2, \ldots, x_n)$$

- Paramêtre : $\theta = t(F)$
- Estimateur : $\hat{\theta} = s(\mathbf{x})$
- biais :

$$\text{biais}_F(\hat{\beta}, \beta) = E_F[s(\mathbf{x})] - t(F)$$

- L'estimation bootstrap du biais :

$$\text{biais}_{\hat{F}} = E_{\hat{F}}[s(\mathbf{x}^*)] - t(\hat{F})$$

  Remarques :
  - $t(\hat{F})$ peut être différent de $\hat{\theta} = s(\mathbf{x})$
  - $\text{biais}_{\hat{F}}$ est l'estimateur "plug-in" de biais_F.
- En pratique :
  - on tire B bootstrap
  - approximation de $\text{biais}_{\hat{F}}$ par :
    - $\hat{\theta}^*(b) = s(\mathbf{x}^*)$
    - $\hat{\theta}^*(.) = \Sigma_{b=1}^{B} \hat{\theta}^*(b)/B$
    - $\text{biais}_B = \hat{\theta}^*(.) - t(\hat{F})$

0.9.2 Exemple : les patchs

UN EXEMPLE : LES PATCHS

Nouveau "patch" médical pour infuser une hormone dans le sang. La compagnie a déjà un ancien "patch".

$$\theta = \frac{E(\text{nouveau}) - E(\text{ancien})}{E(\text{ancien}) - E(\text{placebo})}$$

Acceptation de la Food Drug Administration si : $|\theta| \leq 0.20$ z =vieux-placebo, et y =nouveau-vieux, $\mathbf{x}_i = (z_i, y_i)$

$$F \to \mathbf{x} = (x_1, x_2, \ldots, x_8)$$

$$\theta = t(F) = \frac{E_F(y)}{E_f(z)}$$

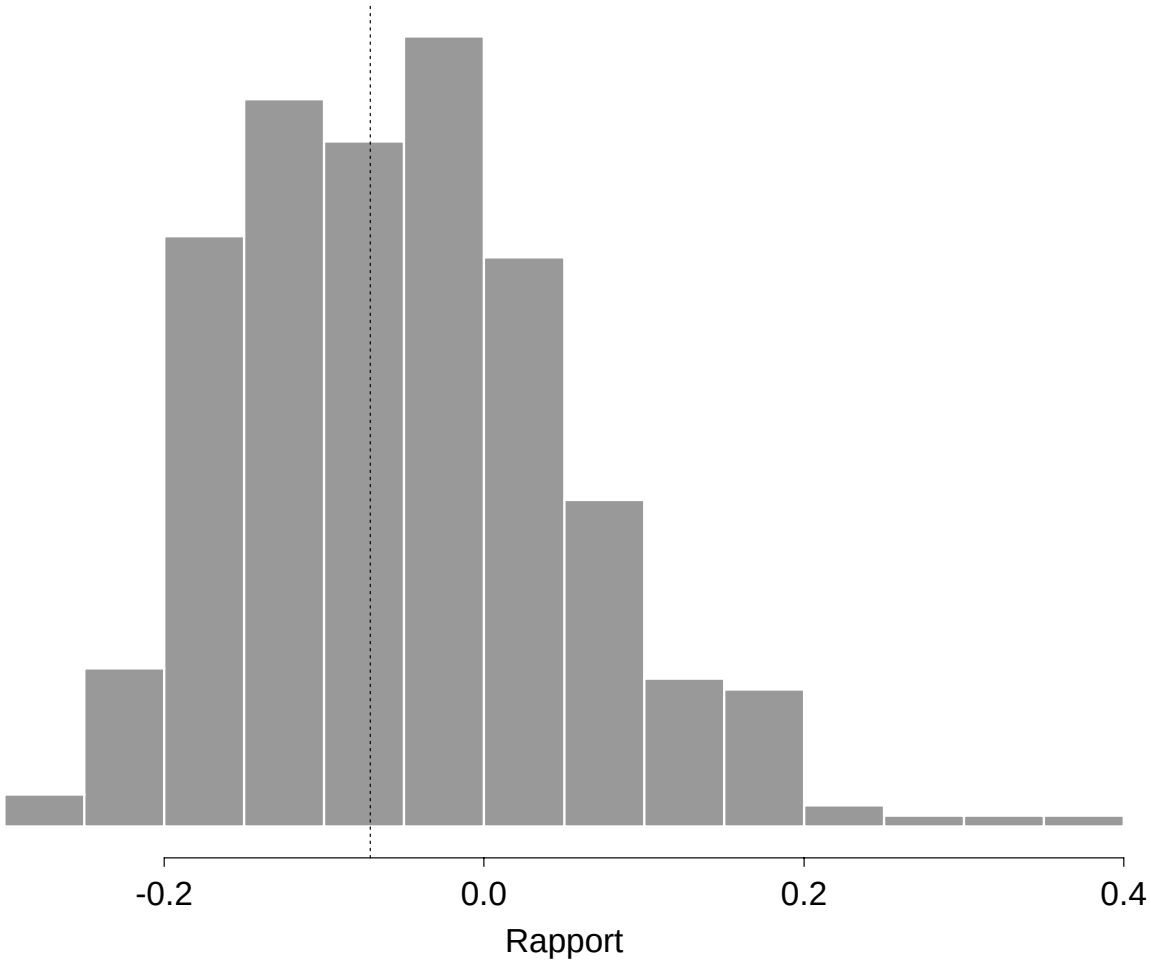
Estimateur "plug-in" :

$$\hat{\theta} = t(\hat{F}) = \frac{\Sigma_{i=1}^{8} y_i/8}{\Sigma_{i=1}^{8} z_i/8} = \frac{-452.3}{6342} = -0.0713$$

0.9.3 Bootstrap

BOOTSTRAP

$$\hat{\theta}^* = \frac{\overline{y}^*}{\overline{z}^*}$$



$$\hat{bias}_{400} = \hat{\theta}^*(.) - \hat{\theta} = -0.0670 - (-0.0713)$$

$$\hat{se}_{400} = 0.105$$

$$\hat{\text{bias}}_{400}/\hat{\text{se}}_{400} = 0.041 < 0.25$$

On en conclut qu'on n'a pas à se soucier du biais.

Mais a-t-on assez de B=400 bootstrap?

0.9.4 Loi des grands nombres

Loi des grands nombres :

$$\begin{aligned} \text{Prob}_{\hat{F}}\{|\hat{\theta}^*(.) - E_{\hat{F}}\{\hat{\theta}^*\}| < 2\frac{\hat{se}_B}{\sqrt{B}}\} \; = \\ \text{Prob}_{\hat{F}}\{|\hat{bias}_{400} - \hat{bias}_{\infty}| < 2\frac{\hat{se}_B}{\sqrt{B}}\} \; \doteq \; 0.95 \end{aligned}$$

en ayant : $\hat{se}_B = 0.105$ et $B = 400$ on a:

$$\text{Prob}_{\hat{F}}\{|\hat{bias}_{400} - \hat{bias}_{\infty}| < 0.0105\} \doteq 0.95$$

L'étendue de cet intervalle est beaucoup plus grand que 0.0043 !

Mais avec une probabilité de 0.95, on a:

$$|\hat{\text{bias}}_{\infty}| < 0.0043 + 0.0105 = 0.0148$$

et le rapport :

$$\hat{\text{bias}}_{\infty}/\hat{\text{se}}_{\infty} < 0.14 < 0.25$$

Donc ce n'est pas grave !

0.10 Le Jacknife

0.10.1 Présentation

LE JACKKNIFE
(couteau suisse)

Echantillon : $\mathbf{x} = (x_1, x_2, \ldots, x_n)$

Estimateur : $\hat{\theta} = s(\mathbf{x})$

estimateur du biais et de l'erreur-standard??

Echantillon Jacknife :

$$\mathbf{x}_{(i)} = (x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$$

Copie Jacknife : $\hat{\theta}_{(i)} = s(\mathbf{x}_{(i)})$

0.10.2 Estimation Jacknife du biais

Estimation Jacknife du biais :

$$\hat{\text{bias}}_{\text{jack}} = (n-1)(\hat{\theta}(.) - \hat{\theta})$$

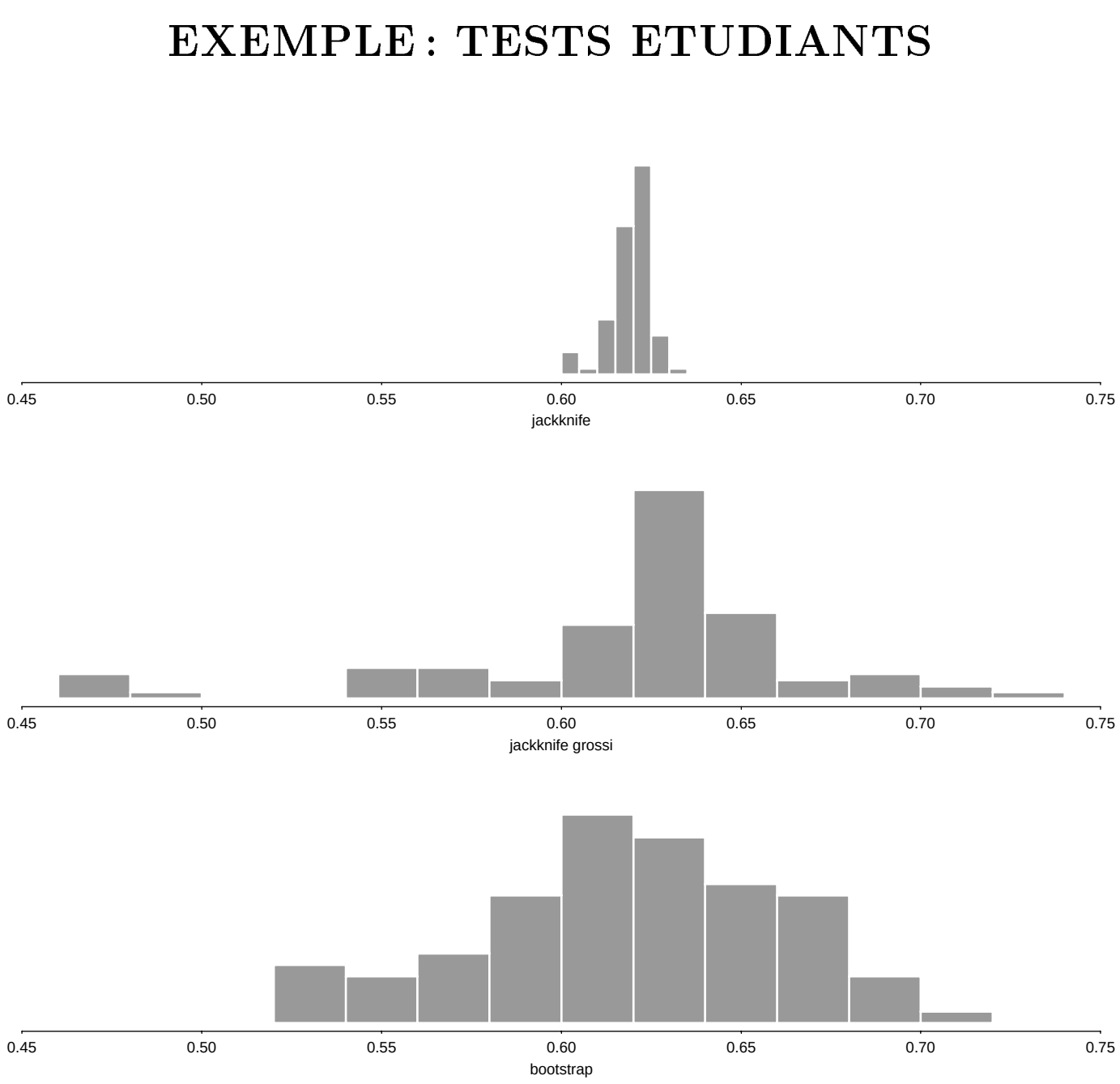
avec :

$$\hat{\theta}(.) = \sum_{i=1}^{n} \hat{\theta}_{(i)}/n$$

Estimation jackknife de l'erreur-standard :

$$\begin{aligned} \hat{\text{se}}_{\text{jack}} &= [\tfrac{n-1}{n} \Sigma(\hat{\theta}_{(i)} - \hat{\theta}(.))^2]^{1/2} \\ &= [\tfrac{1}{n} \Sigma(\sqrt{n-1}(\hat{\theta}_{(i)} - \hat{\theta}(.)))^2]^{1/2} \end{aligned}$$

0.10.3 Exemple : tests étudiants



0.10.4 Relation bootstrap et jacknife

RELATION BOOTSTRAP - JACKKNIFE

– Cas d'une statistique linéaire : $\hat{\theta} = \mu + \frac{1}{n} \Sigma \, \alpha(x_i)$.

Exemple : moyenne $\alpha(x_i) = x_i ; \mu = 0$

Dans ce cas les erreurs-standard sont égales à un facteur $[(n-1)/n)]^{1/2}$ près.

0.10.5 Problèmes du jacknife

PROBLÈMES DU JACKKNIFE

– Jacknife marche bien si la statistique est "lisse" !
Exemple : médiane

– groupe de contrôle des souris :

$$10, 27, 31, 40, 46, 50, 52, 104, 146$$

– Valeurs jackknife :

$$48, 48, 48, 48, 45, 43, 43, 43, 43$$

–

$$\hat{\text{biais}}_{\text{jack}} = 6.68$$

–

$$\hat{\text{biais}}_{100} = 9.58$$

0.10.6 D-jacknife

D-JACKKNIFE

On enlève d observations au lieu d'une seule, avec $n = r.d$
estimation de l'erreur-standard :

$$\hat{\text{bia}}\text{is}_{\text{jack}} = [\frac{r}{C_n^d} \Sigma(\hat{\theta}_{(i)} - \hat{\theta}(.))^2]^{1/2}$$

Pour que ça marche pour la médiane il faut que

$$n^{1/2}/d \to 0 \text{ et } n - d \to \infty$$

en pratique, on prend : $d = \sqrt{n}$

0.11 Intervalles de confiance et Bootstrap

0.11.1 Intervalle "normalisé"

INTERVALLES DE CONFIANCE ET BOOTSTRAP

Intervalle "normalisé" :

- $F \rightarrow \mathbf{x} = (x_1, x_2, \ldots, x_n)$
- Paramêtre : $\theta = t(F)$
- Estimateur : $\hat{\theta} = t(\hat{F})$
- on connaît un estimateur $\hat{\text{se}}$ de l'erreur-standard de $\hat{\theta}$.
- dans beaucoup de circonstances, on applique la loi des grands nombres :

$$Z = \frac{\hat{\theta} - \theta}{\hat{\text{se}}} \sim N(0, 1)$$

En d'autres mots , si $z^{(\alpha)}$ est $100.\alpha$ième percentile de la loi $N(0, 1$:

$$\text{Prob}_F\{z^{(\alpha)} \leq \frac{\hat{\theta} - \theta}{\hat{\text{se}}} \leq z^{(1-\alpha)}\} = 1 - 2\alpha$$

ou comme $z^{(\alpha)} = -z^{(1-\alpha)}$

$$\text{Prob}_F\{\theta \in [\hat{\theta} - z^{(1-\alpha)}\hat{\text{se}}, \hat{\theta} + z^{(1-\alpha)}\hat{\text{se}}]\} = 1 - 2\alpha$$

Exemple : souris :

$$\hat{\theta} = 56.22, \hat{\text{se}} = 13.33$$

Intervalle de confiance “normalisé” à 90 % :

$$56.22 \pm 1.645 \cdot 13.33 = [34.29, 78.15]$$

0.11.2 Intervalle "t-Studentisé"

Intervalle "t-Studentisé" :

- Gosset (1908) :

$$Z = \frac{\hat{\theta} - \theta}{\hat{\text{se}}} \sim t_{n-1}$$

  avec t_{n-1} suit une loi de Student à $n-1$ degrés de liberté.
- Quand n est grand, t_{n-1} ressemble étrangement à $N(0,1)$.
- Meilleur approximation dans le cas de petits échantillons.

Exemple : souris :

$$\hat{\theta} = 56.22, \hat{\text{se}} = 13.33$$

Intervalle de confiance "studentisé" à 90 % :

$$56.22 \pm 1.86 \cdot 13.33 = [31.22, 81.01]$$

0.11.3 Intervalle "t-bootstrapé"

Intervalle "t-bootstrapé" :

- Se libérer des hypothèses de normalité !
- Contruire une table de Z à partir des données .
- Calcul :
  - génération de B bootstraps.
  - Calcul de :
$$Z^*(b) = \frac{\hat{\theta}^*(b) - \hat{\theta}}{\hat{\text{se}}^*(b)}$$
  - le αième percentile de $Z^*(b)$ est estimé par la valeur $\hat{t}^{(\alpha)}$ tel que :
$$\#\{Z^*(b) \leq \hat{t}^{(\alpha)}\}/B = \alpha$$
exemple : B = 1000 .
$$5\ \% \rightarrow 50\text{ième valeur}$$
$$95\ \% \rightarrow 950\text{ième valeur}$$
  - Intervalle de confiance :
$$[\hat{\theta} - \hat{t}^{(1-\alpha)} \cdot \hat{\text{se}}, \hat{\theta} - \hat{t}^{(\alpha)} \cdot \hat{\text{se}}]$$

0.11.4 Exemple : souris

Exemple : souris

Percentile	5%	10%	16%	50%	84%	90%	95%
t_8	-1.86	-1.40	-1.10	0.00	1.10	1.40	1.86
Normal	-1.65	-1.28	-0.99	0.00	0.99	1.28	1.65
t-bootstrapé	-4.53	-2.01	-1.32	-0.025	0.86	1.19	1.53

Intervalle normalisé : [34.29, 78.15]
Intervalle studentisé: [31.22, 81.01]
Intervalle t-bootstrapé : [35.82, 116.74]

Problèmes :

- symétrie de la distribution de la statistique.
- Marche bien pour les statistiques de localisation.
- estimation de $\hat{\text{se}}^*(b)$?? -> double bootstrap !
- ératique si n est petit.

0.11.5 Transformations

TRANSFORMATIONS

On peut obtenir des intervalles de confiance assez farfelus !

Exemple :
coefficient de corrélation des notes des étudiants.
Intervalle de confiance à 98% : $[-0.68, 1.03]$

On transforme pour que les valeurs possibles de l'estimation de la statistique soit réelles.

Exemple :
coefficient de corrélation ρ

$$\Phi = 0.5 \log \frac{1+\rho}{1-\rho}$$

On cherche des transformations qui :

1. normalisent
2. stabilisent la variance

0.11.6 Intervalle des percentiles

Intervalle des percentiles :

- Génération de B bootstraps :

$$\hat{P} \rightarrow \mathbf{x}^* \quad \text{et} \quad \hat{\theta}^* = s(\mathbf{x}^*)$$

- soit $\hat{G}$ la fonction de distribution cumulée des $\hat{\theta}^*$.
- l'intervalle de confiance à $(1-2\alpha)$ basé sur les percentiles est :

$$[\hat{G}^{-1}(\alpha), \hat{G}^{-1}(1-\alpha)]$$

- Si $B.\alpha$ est un entier, l'intervalle de confiance est :

$$[\hat{\theta}_B^{*(\alpha)}, \hat{\theta}_B^{*(1-\alpha)}]$$

Exemple :

$$(x_1, x_2, \ldots, x_{40}) \text{tiré de} N(0,1)$$

Paramêtre : $\theta = e^{\mu}$ où μ est la moyenne. $(\theta = 1)$.
Statistique : $\hat{\theta} = e^{\overline{x}}$

2.5%	5%	10%	16%	50%	84%	90%	95%	97.5%		
0.75	0.82	0.90	0.98		.125	1.61		1.75	1.93	2.07

$$\text{IC}_{\text{perc}} = [0.75, 2.07]$$

$$\text{IC}_{\text{norm}} = [0.59, 1.92]$$

- IC_{norm} marche si la distribution de θ est "Normale".

– Attention : $\mathrm{IC}_{\mathrm{perc}}$ n'arrange pas le biais !

0.11.7 Intervalle accéléré et non-biaisé

Intervalle accéléré et non-biaisé

$$\mathrm{BC}_a : [\hat{\theta}_{\mathrm{bas}}, \hat{\theta}_{\mathrm{haut}}] = [\hat{\theta}_B^{*(\alpha_1)}, \hat{\theta}_B^{*(\alpha_2)}]$$

avec :

$$\begin{aligned}
\alpha_1 &= \Phi(\hat{z}_0 + \frac{\hat{z}_0 + z^{(\alpha)}}{1 - \hat{a}(\hat{z}_0 + z^{(\alpha)})}) \\
\alpha_2 &= \Phi(\hat{z}_0 + \frac{\hat{z}_0 + z^{(1-\alpha)}}{1 - \hat{a}(\hat{z}_0 + z^{(1-\alpha)})})
\end{aligned}$$

$$\Phi(\cdot) = \text{Fonction de distribution cumulée de} N(0,1)$$

$$z^{(\alpha)} = 100\alpha\text{ième percentile de} N(0,1)$$

$$\begin{aligned}
\hat{z}_0 &= \Phi^{-1}(\frac{\#\{\hat{\theta}^*(b) < \hat{\theta}\}}{B}) \\
\hat{a} &= \frac{\sum_{i=1}^n (\hat{\theta}_{(.)} - \hat{\theta}_{(i)})^3}{6\{\sum_{i=1}^n (\hat{\theta}_{(.)} - \hat{\theta}_{(i)})^2\}_{3/2}}
\end{aligned}$$

1. si $\hat{a}$ et z_0 sont nuls alors $\mathrm{BC}_a = \mathrm{ICV}_{\mathrm{perc}}$.
2. z_0 corrige le biais. Il mesure le biais médian de $\hat{\theta}^*$.
3. $\hat{a}$ est l'accélération. L'approximation $\hat{\theta}\ N(\theta, \mathrm{se}^2)$ suppose que la variance ne dépend pas de θ. $\hat{a}$ corrige ce possible défaut.

1. Comme l'$\mathrm{IC}_{\mathrm{perc}}$, le BC_a conserve les transformations.

   exemple : le BC_a de $\sqrt{\theta}$ est calculé en prenant les racines carrées de $[\hat{\theta}_{\mathrm{bas}}, \hat{\theta}_{\mathrm{haut}}]$

$$\mathrm{BC}_a(\sqrt{\theta}) = [\sqrt{\hat{\theta}_{\mathrm{bas}}}, \sqrt{\hat{\theta}_{\mathrm{haut}}}]$$

2. BC_a est plus précis que les autres.

   Pour BC_a : $\mathrm{Prob}\{\theta < \hat{\theta}_{\mathrm{bas}}\} = \alpha + \frac{c_{\mathrm{bas}}}{n}$

   Pour $\mathrm{IC}_{\mathrm{perc}}, \mathrm{IC}_{\mathrm{norm}}$: $\mathrm{Prob}\{\theta < \hat{\theta}_{\mathrm{bas}}\} = \alpha + \frac{c_{\mathrm{bas}}}{\sqrt{n}}$

0.12 Tests de permutation

0.12.1 Historique

TESTS DE PERMUTATION

– Fisher 1930. (t de Student).

– Problèmes d'hypothèses mathématiques.

– Exemple : 2 échantillons indépendants.

–

$$\begin{array}{rcl} F & \longrightarrow & \mathbf{z} = (z_1, z_2, \ldots, z_m) \\ & & \text{indépendamment de} \\ G & \longrightarrow & \mathbf{y} = (y_1, y_2, \ldots, y_n) \end{array}$$

– Hypothèse nulle : $\mathrm{H}_0 : F = G$

– H_0 c'est l'avocat du diable.

– exemple des souris : un souhait $\hat{\theta} = \overline{z} - \overline{y} > 0$

– Niveau de signification :

$$\mathrm{ASL} = \mathrm{Prob}_{\mathrm{H}_0}\{\hat{\theta}^* \geq \hat{\theta}\}$$

– $\hat{\theta}^*$ est une variable aléatoire qui suit la loi de distribution de $\hat{\theta}$ si H_0 est vraie

– Plus ASL est petit, plus on a de chances que H_0 soit vraie.

0.12.2 Exemple : les souris

EXEMPLE "LES SOURIS"

– Hypothèses : F et G sont des distributions normales

$$F = N(\mu_T, \sigma^2) G = N(\mu_C, \sigma^2)$$

– Hypothèse nulle :

$$\mathrm{H}_0 : \mu_T = \mu_C$$

– ou

$$\mathrm{H}_0 : \hat{\theta} \sim N(0, \sigma^2(\frac{1}{n} + \frac{1}{m})$$

–

$$\begin{aligned} \mathrm{ASL} &= \mathrm{Prob}\{Z > \frac{\hat{\theta}}{\sigma\sqrt{1/n+1/m}}\} \\ &= 1 - \Phi(\frac{\hat{\theta}}{\sigma\sqrt{1/n+1/m}}) \end{aligned}$$

– Mais σ est inconnu Estimation :

$$\overline{\sigma} = \{\sum_{i=1}^{n}[(z_i - \overline{z})^2 + \sum_{j=1}^{m}(y_j - \overline{y})^2]/[n+m-2]\}^{1/2}$$

–

$$\mathrm{ASL} = 1 - \Phi(\frac{30.63}{54.21\sqrt{1/9+1/7}}) = 0.131$$

– t de Student :

$$\mathrm{ASL} = \mathrm{Prob}\{t_{14} \leq \frac{30.63}{54.21\sqrt{1/9+1/7}}\} = 0.141$$

– On ne peut pas rejeter H_0

0.12.3 L'idée

L'IDÉE

On peut ranger les valeurs z_i et y_j par ordre croissant et indiquer pour chaque valeur à quel groupe elle appartient.

valeur :	10	16	23	27	31	38	40	46
groupe :	y	z	z	y	y	z	y	y
valeur :	50	52	94	99	104	141	146	197
groupe :	y	y	z	z	y	z	y	z

$$\begin{aligned} \mathbf{v} &: \text{vecteur des N valeurs} \\ \mathbf{g} &: \text{vecteur des groupes} \\ N &= n + m \end{aligned}$$

la donnée du couple (**g**,**v**) est identique à donner le couple (**z**,**y**)

Nombre de vecteurs **g** possibles : C_N^m

Sous H_0, le vecteur $\mathbf{g}$ a la probabilité $1/C_N^m$ de valoir chacune de ces valeurs possibles

La statistique $\hat{\theta}$ est une fonction de $(\mathbf{g},\mathbf{v})$:

$$\hat{\theta} = S((\mathbf{g},\mathbf{v}))$$

Pour chacune des C_N^m vecteurs possibles de $\mathbf{g}$ on peut calculer :

$$\hat{\theta}^* = \hat{\theta}(\mathbf{g}^*) = S(\mathbf{g}^*, \mathbf{v})$$

Le niveau de signification est :

$$\begin{aligned} \mathrm{ASL}_{\mathrm{perm}} &= \mathrm{Prob}_{\mathrm{perm}}\{\hat{\theta}^* \leq \hat{\theta}\} \\ &= \#\{\hat{\theta}^* \leq \hat{\theta}\}/C_N^m \end{aligned}$$

0.12.4 Calcul de la statistique par test de permutation

Calcul de la statistique par test de permutation

1. Choisir B vecteurs indépendants $\mathbf{g}^*(1), \mathbf{g}^*(2), \ldots, \mathbf{g}^*(B)$ parmi les C_N^m vecteurs possibles.
2. Calculer :
$$\hat{\theta}(b) = S(\mathbf{g}^*, \mathbf{v})$$
3. Calculer l'approximation :
$$\hat{\text{ASL}}_{\text{perm}} = \#\{\hat{\theta}^* \leq \hat{\theta}\}/B$$

NOMBRE DE PERMUTATIONS

ASL_{perm} :	0.5	0.25	0.10	0.005	0.025
B :	100	300	900	2000	4000

Exemple des souris : $\text{ASL}_{\text{perm}} = 0.132$

- Aucune hypothèse n'est faite sur F et G.
- On peut remplacer la différence des moyennes par la différence d'une autre tendance centrale.
- Dans la pratique on tire N valeurs uniforme entre 0 et 1, puis les n plus petites valeurs sont dites appartenir au pre-

mier groupe et les autres aux deuxième groupe.

0.12.5 Un autre exemple

TESTS DE PERMUTATION : UN AUTRE EXEMPLE

PROBLÈME :
COMPARER 2 MARQUES D'AIGUILLES UTILISÉES POUR DES PRÉLÈVEMENTS SANGUINS.

Marque	Nombre de prélèvements	Nombre d'infections
A	40	4
B	30	0

QUESTION :
PEUT-ON ATTRIBUER LA DIFFÉRENCE DE 4 AU HASARD ?

TEST DE PERMUTATION :

1. faire B fois :
   (a) générer 40 nombres entiers au hasard entre 1 et 70.
   (b) compter le nombre de fois g_1 où les entiers 1 à 4 (supposés être infectés) apparaissent.
   (c) générer 30 nombres entiers au hasard entre 1 et 70.
   (d) compter le nombre de fois g_2 où les entiers 1 à 4 apparaissent.
   (e) calculer la différence $\hat{\delta}^* = g_1 - g_2$.
2. calculer :

$$\hat{\text{ASL}}_{\text{perm}} = \#\{\hat{\delta}^* \geq 4\}/B$$

DANS NOTRE EXEMPLE :

$$B = 3000 \text{ et } \hat{\text{ASL}}_{\text{perm}} = 0.043$$

Bibliographie

[1] P. Diaconis et B. Efron. *Méthodes de calculs statistiques intensifs sur ordinateurs.* dans *Le calcul intensif.* Bibliothèque Pour La Science. (1989).

[2] B. Efron. *An Introduction to the Bootstrap.* Chapmann & Hall . (1993).

[3] F. Mosteller et J. W. Tukey. *Data Analysis and Regression.* Addison-Wesley (1977).