MUSÉUM NATIONAL D'HISTOIRE NATURELLE

SEMIN-R

Rééchantillonnage sous R : bootstrap et jackknife

Loïc PONGER

MNHN
USM 503 Régulation et Dynamique des Génomes
ponger@mnhn.fr

Rééchantillonnage sous R

bootstrap et jackknife

Ponger Loic

MNHN-USM503
43 ru Cuvier
75005 Paris

6 mai 2008

MUSÉUM NATIONAL D'HISTOIRE NATURELLE

Bootstrap et Jackknife

Méthodes statistiques basées sur un rééchantillonnage des données à partir d'une distribution estimée de la population réelle (inconnue). Cette distribution est "souvent" construire à partir de l'échantillon observé.

Objectifs

- $\Rightarrow$ estimer un intervalle de confiance autour d'un paramètre,
- $\Rightarrow$ construire un test d'hypothèse,
- $\Rightarrow$ identifier des valeurs abérrantes (jackknife).

Propriétés

- $\Rightarrow$ simple, paramétrique *vs.* non-paramétrique,
- $\Rightarrow$ "computer intensive"

MUSÉUM NATIONAL D'HISTOIRE NATURELLE

MUSÉUM NATIONAL D'HISTOIRE NATURELLE

MUSÉUM NATIONAL D'HISTOIRE NATURELLE

Le bootstrap

On calcule une statistique (moyenne, médiane, r^2, ...) à partir d'un échantillon de n individus issues de la pop. d'étude. La distribution de la population est inconnue ou non-normale. Le bootstrap permet de construire une distribution proche de la distribution inconnue.

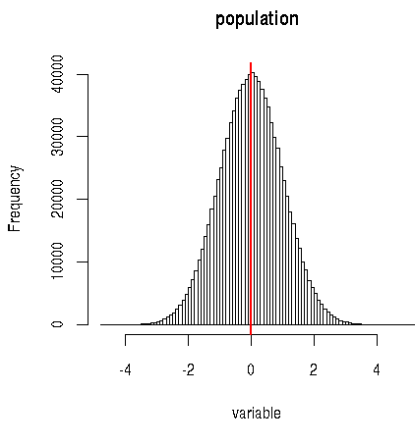
Algorithme

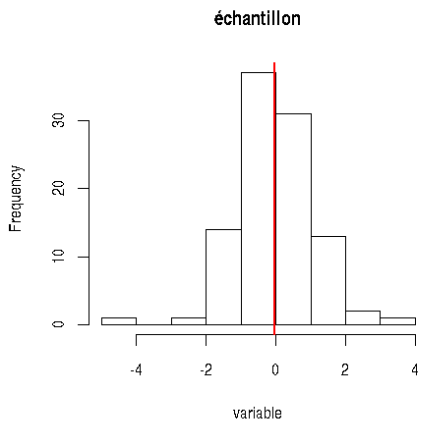
- Calcul de la statistique observée
- Pour i allant de 1 à n_b
  - échantillonnage aléatoire avec remise de n individus parmis l'échantillon observé
  - calcul de la statistique sur le i^{ieme} échantillon de bootstrap
- Calcul la fonction de distribution à partir des n_b valeurs de la statistique $\Rightarrow$ intervalle de confiance ou autre

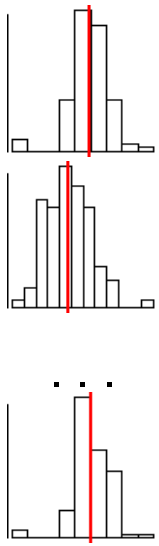
Hyp de travail : les individus sont indépendants et identiquement distribués (i.i.d.)

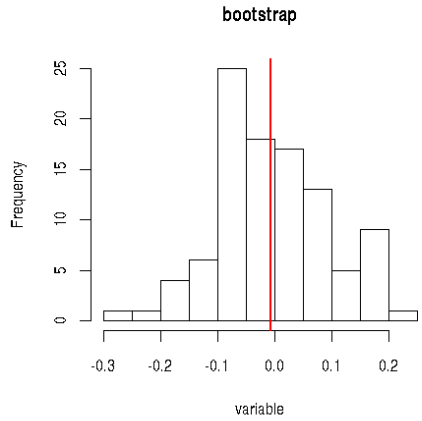
MUSÉUM NATIONAL D'HISTOIRE NATURELLE

Le bootstrap : principe









population étudiée (μ)

échantillon observé ($\bar{x}$)

B échantillons de bootstrap ($\bar{x^*}$)

distribution des $\bar{x^*}$ (moyenne des $\bar{x^*}$)

MUSÉUM NATIONAL D'HISTOIRE NATURELLE

Le bootstrap : manuellement

Algorithme

```
###   Les données
goudron=c(0.45,0.77,1.07,1.03,1.34,1.14,1.15,0.9,0.55,1.15)
nicotine=c(11,13,14,15,17,18,14.5,13.5,8.5,16.5)
tabac=cbind(goudron,nicotine)
l=length(nicotine)

tabac
   goudron nicotine
 [1,]    0.45     11.0
 [2,]    0.77     13.0
 [3,]    1.07     14.0
 [4,]    1.03     15.0
 [5,]    1.34     17.0
 [6,]    1.14     18.0
 [7,]    1.15     14.5
 [8,]    0.90     13.5
 [9,]    0.55      8.5
[10,]    1.15     16.5
```

Example issu de Tomassone R., Charles-Bajard S. et Bellanger L. (1998)

MUSÉUM NATIONAL D'HISTOIRE NATURELLE

Le bootstrap : manuellement

Algorithme

```
###   La fonction pour calculer la statistique à
###         partir des indices
functionCor=function(val,  ii){
cor(    val[ii,1]    ,   val[ ii,2]     )
}

### Corrélation observée
functionCor(tabac,1 :1)
0.8901696
```

Example issu de Tomassone R., Charles-Bajard S. et Bellanger L. (1998)

MUSÉUM NATIONAL D'HISTOIRE NATURELLE

Le bootstrap : manuellement

Algorithme

```
### Bootstrap
NB=1000
bootCor=vector(length=NB)
for (ind in 1 :NB) {
     bootIndice=sample(1 :l, replace=T)
 bootCor[ind]=functionCor(tabac, bootIndice)
}

### calcul de l'IC à 95%
quantile(bootCor, c(0.025, 0.975))
    2.5%     97.5%
0.7175133 0.9731658
```

Example issu de Tomassone R., Charles-Bajard S. et Bellanger L. (1998)

MUSÉUM NATIONAL D'HISTOIRE NATURELLE

R : biliothèques de bootstrap

Deux bibliothèques

bootstrap permet également de faire du jackknife, (d'après Efron et Tibshirani, 1994), maintenue par K. Halvorsen

boot plus de fonctionnalités (d'après Davison et Hinkley, 1997), maintenue par B. Ripley

MUSÉUM NATIONAL D'HISTOIRE NATURELLE

MUSÉUM NATIONAL D'HISTOIRE NATURELLE

R : le bootstrap facile

Algorithme

```
library(boot)

###   Les donnéesgoudron=c(0.45,0.77,1.07,1.03,1.34,1.14,1.15,0.9,0.55,1.15)
nicotine=c(11,13,14,15,17,18,14.5,13.5,8.5,16.5)
tabac=cbind(goudron,nicotine)
l=length(nicotine)

###  Définition de la fonction pour le calcul de la statistique

myCor <- function(data, indice){
cor(data[indice,1],data[indice,2])
}

b=boot(tabac, statistic=myCor, R=1000, sim="ordinary", stype="i")
```

On utilise ici la fonction boot de la bibliothèque boot. On pourrait également utiliser la fonction bootstrap (bibliothèque bootstrap) mais l'utilisation de la fonction pour le calcul de la statistique est différent.

MUSÉUM NATIONAL D'HISTOIRE NATURELLE

R : le bootstrap facile

Algorithme

```
summary(b)
          Length Class  Mode
t0           1   -none- numeric
t         1000   -none- numeric
R            1   -none- numeric
data        20   -none- numeric
seed       626   -none- numeric
statistic    1   -none- function
sim          1   -none- character
call         6   -none- call
stype        1   -none- character
strata      10   -none- numeric
weights     10   -none- numeric
```

On utilise ici la fonction boot de la bibliothèque boot. On pourrait également utiliser la fonction bootstrap (bibliothèque bootstrap) mais l'utilisation de la fonction pour le calcul de la statistique est différent.

MUSÉUM NATIONAL D'HISTOIRE NATURELLE

R : le bootstrap facile

Algorithme

```
summary(b$t)
       V1
 Min.   :0.4840
 1st Qu. :0.8774
 Median :0.9061
 Mean   :0.8947
 3rd Qu. :0.9302
 Max.   :0.9939
```

On utilise ici la fonction boot de la bibliothèque boot. On pourrait également utiliser la fonction bootstrap (bibliothèque bootstrap) mais l'utilisation de la fonction pour le calcul de la statistique est différent.

MUSÉUM NATIONAL D'HISTOIRE NATURELLE

R : la fonction boot

boot(data, statistic, R, sim="ordinary", stype="i", ...)

les arguments

data	vecteur ou table (observations sur les lignes)
statistic	fonction retournant la valeur de la statistique
R	nombre de réplicats
sim	type de bootstrap (ordinary, parametric, ...)
stype	dépend de la fonction utilisée

- indices : i, numéros des lignes sélectionnées (certains numéros apparaissent plusieurs fois)
- fréquences (absolues) : f, un entier correspondant au nombre de sélection de chaque ligne (somme est égale au nombre d'observations initiales)
- poids : w, un réel compris entre 0 et 1 (somme est égale à 1)

R : la fonction boot

les sorties

- `t0` la statistique calculée sur les données
- `t` les valeurs de la statistiques calculées sur les réplicats (R valeurs)

...

MUSÉUM NATIONAL D'HISTOIRE NATURELLE

R : la fonction boot

la fonction statistique

Elle est calculée par une fonction prenant au moins 2 arguments :

- **le premier** les données
- **le second** un vecteur avec les indices/poids/fréquences des valeurs sélectionnées (par la fonction boot). Si les données sont sous la forme de tableau, les indices représentent les numéros de lignes.
- **option** des arguments qui seront passés via boot

Une fonction calculant la moyenne à partir des indices :

```
statistic=function(valeurs,indices){
sum(valeurs[indices])/length(indices)
}
boot(data, stat=statistic, R=1000, stype="i")
```

MUSÉUM NATIONAL D'HISTOIRE NATURELLE

R : la fonction boot

la fonction statistique

Elle est calculée par une fonction prenant au moins 2 arguments :

- **le premier** les données
- **le second** un vecteur avec les indices/poids/fréquences des valeurs sélectionnées (par la fonction boot). Si les données sont sous la forme de tableau, les indices représentent les numéros de lignes.
- **option** des arguments qui seront passés via boot

Une fonction calculant la moyenne à partir des fréquences :

```
statistic=function(valeurs,frequences){
sum(valeurs*frequences)/sum(frequences))
}
boot(data, stat=statistic, R=1000, stype="f")
```

MUSÉUM NATIONAL D'HISTOIRE NATURELLE

R : la fonction boot

la fonction statistique

Elle est calculée par une fonction prenant au moins 2 arguments :

- **le premier** les données
- **le second** un vecteur avec les indices/poids/fréquences des valeurs sélectionnées (par la fonction boot). Si les données sont sous la forme de tableau, les indices représentent les numéros de lignes.
- **option** des arguments qui seront passés via boot

Une fonction calculant une moyenne à partir de poids (sens ?) :

```
statistic=function(valeurs,poids){
sum(valeurs*poids)/1
}
boot(data, stat=statistic, R=1000, stype="w")
```

MUSÉUM NATIONAL D'HISTOIRE NATURELLE

R : la fonction boot

la fonction statistique

Elle est calculée par une fonction prenant au moins 2 arguments :

- **le premier** les données
- **le second** un vecteur avec les indices/poids/fréquences des valeurs sélectionnées (par la fonction boot). Si les données sont sous la forme de tableau, les indices représentent les numéros de lignes.
- **option** des arguments qui seront passés via boot

Une fonction calculant la moyenne en utilisant mean (et un troisième arg.) :

```
statistic=function(valeurs, indices, t){
mean(valeurs[indices], trim=t)
}
boot(data, stat=statistic, R=1000, stype="i", t=0.25)
```

MUSÉUM NATIONAL D'HISTOIRE NATURELLE

R : la fonction boot

les différents type de ré-échantillonnage

ordinary bootstrap Chaque échantillon de bootstrap est issu d'un tirage alétoire. Les échantillons sont indépendants des précédents.

parametric bootstrap Les n_b échantillons sont construits à partir d'une distribution théorique dont les paramètres de position et variabilité sont estimés à partir de l'échantillon.

balanced bootstrap Chaque valeur apparait n_b fois parmis tous les n_b rééchantillonnages. Augmente la précision de SE.

antithetic resampling Chaque échantillon de bootstrap est construit en utilisant les couples (i, n-i). Réduit la variance de la statistique.

permutation Chaque échantillon de bootstrap est composé par une permutation aléatoire des n valeurs.

MUSÉUM NATIONAL D'HISTOIRE NATURELLE

R : la fonction boot.ci

les 5 types d'estimation des CI

normal Utilisation d'une loi normale dont les paramètres (moyenne et variance) sont estimés à partir de la distribution empirique.

basic Utilisation la distribution des $\hat{\theta}_i - \hat{\theta}$.

student Utilisation de la distribution des $t_i = \frac{\hat{\theta}_i - \hat{\theta}}{s_{\hat{\theta}_i}}$.

percent Utilisation de la distribution enmpirique des $\hat{\theta}_i$

bca Bias Corrected and Accelerated. Transformation de la distribution empirique en une loi normale centrée réduite. Correction pour l'asymmétrie.

Voir : Carpenter and Bithell (2000) (papier relativement simple et clair)

MUSÉUM NATIONAL D'HISTOIRE NATURELLE

MUSÉUM NATIONAL D'HISTOIRE NATURELLE

R : les autres fonctions

censboot	fonction de bootstrap pour données "right-censored"
tsboot	fonction de bootstrap pour des séries temporelles (découpage en blocs)
tilt.boot	fonction de bootstrap basée sur un "importance resampling"
empinf	retourne l'influence empirique des valeurs sur la statistique étudiée à partir d'un objet boot
boot.array	retourne un tableau $n * n_b$ avec les fréq. de chaque individu à partir d'un objet boot
jack.after.boot	effectue un jackknife sur un objet boot.

MUSÉUM NATIONAL D'HISTOIRE NATURELLE

1. Bootstrap
   - présentation
   - la fonction boot
   - la fonction boot.ci
   - les autres fonctions

2. Jackknife
   - présentation
   - la fonction jackknife
   - la fonction jack.after.boot

MUSÉUM NATIONAL D'HISTOIRE NATURELLE

1. Bootstrap
   - présentation
   - la fonction boot
   - la fonction boot.ci
   - les autres fonctions

2. Jackknife
   - présentation
   - la fonction jackknife
   - la fonction jack.after.boot

MUSÉUM NATIONAL D'HISTOIRE NATURELLE

Le jackknife

On calcule une statistique (moyenne, médiane, r^2, ...) à partir d'un échantillon de n individus issues de la pop. d'étude. La distribution de la population est inconnue ou non-normale. Le jackknife permet de construire une distribution proche de la distribution inconnue.

Algorithme

- Calcul de la statistique observée
- Pour i allant de 1 à n
  - échantillonnage de tous les individus, sauf le i^{ieme}
  - calcul de la statistique sur le i^{ieme} échantillon de bootstrap
- Calcul la fonction de distribution à partir des n valeurs de la statistique $\Rightarrow$ intervalle de confiance ou autre
- Analyse des n valeurs de la statistique à la recherche de valeurs abérrantes

Le Jackknife : manuellement

Algorithme

```
###   Les données
goudron=c(0.45,0.77,1.07,1.03,1.34,1.14,1.15,0.9,0.55,1.15)
nicotine=c(11,13,14,15,17,18,14.5,13.5,8.5,16.5)
tabac=cbind(goudron,nicotine)
l=length(nicotine)

tabac
   goudron nicotine
 [1,]    0.45     11.0
 [2,]    0.77     13.0
 [3,]    1.07     14.0
 [4,]    1.03     15.0
 [5,]    1.34     17.0
 [6,]    1.14     18.0
 [7,]    1.15     14.5
 [8,]    0.90     13.5
 [9,]    0.55      8.5
[10,]    1.15     16.5
```

MUSÉUM NATIONAL D'HISTOIRE NATURELLE

Le Jackknife : manuellement

Algorithme

```
###   La fonction pour calculer la statistique à
###         partir des indices
functionCor=function(val,  removed){
cor(    val[ -removed,1]     ,    val[  -removed,2]     )
}

### Corrélation observée
functionCor(tabac,1 :1)
0.8901696
```

MUSÉUM NATIONAL D'HISTOIRE NATURELLE

Le Jackknife : manuellement

Algorithme

```
### Jackknife
jackCor=vector(length=l)
for (ind in 1 :l) {
 jackCor[ind]=functionCor(tabac, ind)
}

### calcul de l'IC à 95%
quantile(jackCor, c(0.025, 0.975))
    2.5%     97.5%
0.8722099 0.9126298
```

MUSÉUM NATIONAL D'HISTOIRE NATURELLE

MUSÉUM NATIONAL D'HISTOIRE NATURELLE

R : le jackknife facile

Algorithme

```
library(bootstrap)

###   Les donnéesgoudron=c(0.45,0.77,1.07,1.03,1.34,1.14,1.15,0.9,0.55,1.15)
nicotine=c(11,13,14,15,17,18,14.5,13.5,8.5,16.5)
tabac=cbind(goudron,nicotine)
l=length(nicotine)

###  Définition de la fonction pour le calcul de la statistique

myCor <- function(indice, data){
cor(data[indice,1],data[indice,2])
}
 j=jackknife(  1 :l,    myCor,      data=tabac)
```

MUSÉUM NATIONAL D'HISTOIRE NATURELLE

R : le jackknife facile

Algorithme

```
summary(j)
            Length Class  Mode
jack.se      1     -none- numeric
jack.bias    1     -none- numeric
jack.values 10     -none- numeric
call         4     -none- call


summary(j$jack.values)
      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.8710  0.8848  0.8902  0.8923  0.9012  0.9146
```

MUSÉUM NATIONAL D'HISTOIRE NATURELLE

R : La fonction jackknife

Utiliser le package bootstrap :

les arguments

x	le vecteur de valeurs
theta	la fonction calculant la statistique La fonction theta reçoit le vecteur x privé du i^{ieme} élément comme argument.

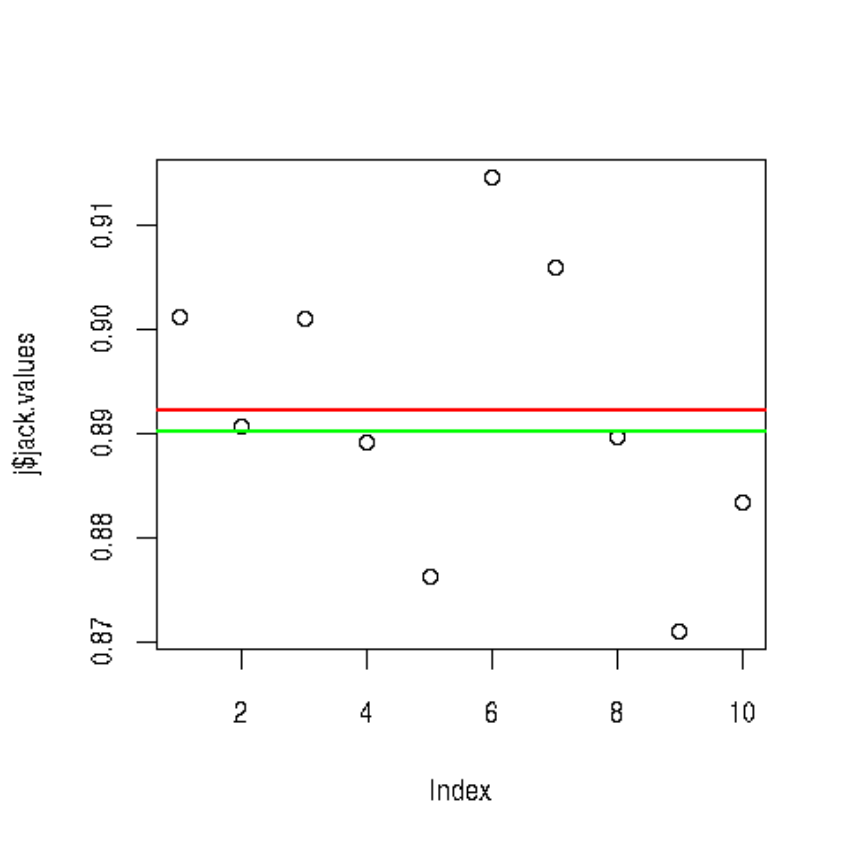
Attention : la fonction theta du package bootstrap ne fonctionne pas comme celle du package boot.

les sorties

jack.se	: déviation standard de la statistique
jack.bias	: biais de la statistique
jack.values	: les n valeurs obtenues en retirant successivement les n individus

Le jackknife pour identifier les valeurs abérrantes

```
 j=jackknife(  1:l,    myCor,      data=tabac)}\only<2>{
plot(j$jack.values)
```



R : comment écrire la fonction theta

Données simples : on peut fournir la liste des valeurs x . La fonction jackknife "élimine" le i^{ieme} individu de x et lance la fonction theta avec les valeur restantes.

jackknife sur les valeurs x_i

```
x <- rnorm(20)
theta <- function(x)mean(x)
results <- jackknife(x,theta)
```

Données complexes : il faut stocker les données dans un tableau et fournir la liste des lignes à traiter. La fonction jackknife "élimine" la i^{ieme} ligne de x et lance la fonction theta avec les numéros de lignes restantes. Le tableau doit être passé en argument supplémentaire.

jackknife sur les indices i

```
xdata <- matrix(rnorm(30),ncol=2)
  n <- 15
  theta <- function(x,xdata) cor(xdata[x,1],xdata[x,2])
  results <- jackknife(1 :n,theta,xdata)
```

1. Bootstrap
   - présentation
   - la fonction boot
   - la fonction boot.ci
   - les autres fonctions

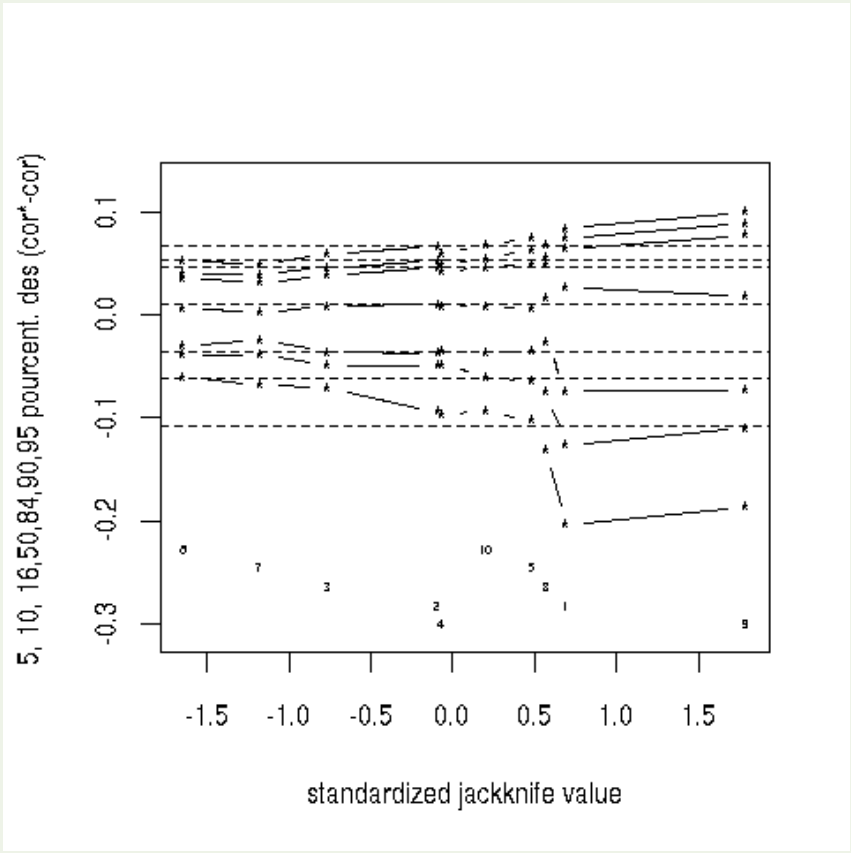
2. Jackknife
   - présentation
   - la fonction jackknife
   - la fonction jack.after.boot

MUSÉUM NATIONAL D'HISTOIRE NATURELLE

Le jackknife après le bootstrap

Objectif : identifier des valeurs abérrantes parmis les n_b bootstraps

```
library(boot)
jack.after.boot(b)
```



Références

approche générale P. Hall (2003) A Short Prehistory of the Bootstrap. Statist. Sci., 18 :158-167.

approche générale P. M. Dixon (2002) Bootstrap resampling. Encyclopedia of Environmetrics, 1 :212-220

calcul des CI J. carpenter and J. Bithell (2000) Bootstrap confidence intervals : when, which, what ? A practical guide for medical statisticians. Statistics in Medicine. 19 :1141-1164.

exemples http ://www.stat.umn.edu/geyer/old03/5601/examp

MUSÉUM NATIONAL D'HISTOIRE NATURELLE