Introduction à l'approche bootstrap

Irène Buvat
U494 INSERM

buvat@imed.jussieu.fr

25 septembre 2000

Plan du cours

- Qu'est-ce que le bootstrap ?
- Bootstrap pour l'estimation d'erreurs standard
- Bootstrap de données structurées
- Bootstrap pour l'estimation de biais
- Bootstrap et jackknife
- Bootstrap pour la construction d'intervalles de confiance
- Bootstrap et tests d'hypothèses
- Bilan
- Référence

Qu'est-ce que le bootstrap ?

• Technique permettant d'effectuer de l'inférence statistique

• Technique récente (1979) car reposant sur l'usage de calculateurs puissants

• Technique reposant sur la simulation de données à partir d'un nombre limité d'observations

• Technique destinée à faciliter l'inférence dans les situations complexes où les méthodes analytiques ne suffisent pas

to pull oneself up by one's bootstrap

= se tirer d'un mauvais pas

Problématique : exemple d'inférence statistique

- La différence entre deux valeurs moyenne est-elle statistiquement significative ?

durée de survie

groupe 1 (placébo)	groupe 2 (traitement)
$n_1 = 9$ mesures	$n_2 = 7$ mesures
52, 10, 40, 104, 50, 27, 146, 31, 46	94, 38, 23, 197, 99, 16, 141
moyenne $m_1 = 56.22$	moyenne $m_2 = 86.86$
erreur standard $se_1 = \sqrt{var_1/n_1} = 14.14$	erreur standard $se_2 = \sqrt{var_2/n_2} = 25.24$

différence des moyennes = 30.63

erreur standard associée à la différence

$$se = \sqrt{se_1^2 + se_2^2} = \sqrt{14.14^2 + 25.24^2} = 28.93$$

$$\frac{m_1 - m_2}{se} = 1.05$$

non significatif

pas besoin de bootstrap !

Problématique : intérêt du bootstrap

- La différence entre deux valeurs médianes est-elle statistiquement significative ?

durée de survie	
groupe 1 (placébo) $n_1 = 9$ mesures	groupe 2 (traitement) $n_2 = 7$ mesures
52, 10, 40, 104, 50, 27, 146, 31, 46	94, 38, 23, 197, 99, 16, 141
médiane $\theta_1 = 46$ erreur standard ?	moyenne $\theta_2 = 94$ erreur standard ?

différence des moyennes = 48

erreur standard associée à la différence ?

différence significative ?

pas de formule analytique simple pour estimer la fiabilité des grandeurs autres que les valeurs moyennes

➥ intérêt du bootstrap

Bootstrap pour l'estimation d'une erreur standard

1 échantillon observé $\mathbf{x} = (x_1, x_2, \ldots, x_N)$ et 1 statistique d'intérêt $s(\mathbf{x})$: moyenne, médiane,...

↓

B échantillons bootstrap

↓

	calcul de la statistique d'intérêt	réplications bootstrap de s
$\mathbf{x}^{*1} = (x_1^*, x_2^*, \ldots, x_N^*)$	→	$s(\mathbf{x}^{*1})$
$\mathbf{x}^{*b} = (x_1^*, x_2^*, \ldots, x_N^*)$	→	$s(\mathbf{x}^{*b})$
…		…
$\mathbf{x}^{*B} = (x_1^*, x_2^*, \ldots, x_N^*)$	→	$s(\mathbf{x}^{*B})$

➥ estimée bootstrap de l'erreur standard = écart-type des réplications bootstrap

$$\sqrt{\frac{\sum_b [s(\mathbf{x}^{*b}) - s^*]^2}{B-1}}$$

avec $s^* = \sum_b s(\mathbf{x}^{*b})/B$

Calcul d'un échantillon bootstrap

1 échantillon observé de N valeurs

$\mathbf{x}$ = (50 , 53 , 58 , 80 , 75 , 69 , 77 , 44 , 63 , 73)

|

1 échantillon bootstrap :
1 tirage aléatoire de N valeurs
parmi l'échantillon original, avec remise

↓

$\mathbf{x}^{*1}$ = (69 , 53 , 80 , 69 , 73 , 53 , 44 , 58 , 75 , 53)

- 1 échantillon bootstrap :
  - ➥ autant de valeurs que dans l'échantillon original
  - ➥ valeurs issues de l'échantillon original, mais avec des fréquences potentiellement différentes

Exemple : erreur standard de la moyenne

durée de survie

groupe 1 (placebo)
$n_1 = 9$ mesures

$\mathbf{x}$ = (52, 10, 40, 104, 50, 27, 146, 31, 46)
statistique d'intérêt : moyenne $m_1 = 56.22$

B échantillons bootstrap

		réplications bootstrap de la moyenne
$\mathbf{x}^{*1}$=(50, 10, 40, 50, 46, 10,146, 40, 50)	calcul de la moyenne →	49.11
$\mathbf{x}^{*b}$ =(10, 52, 104, 40, 104, 46, 50, 146, 27)	→	64.33
…		…
$\mathbf{x}^{*B}$ =(146, 31, 31, 10, 27, 40, 104, 46, 50)	→	53.89

➥ estimée bootstrap de l'erreur standard
= écart-type des réplications bootstrap de la moyenne

$$SE(m_1) = \sqrt{\frac{\sum_b [m_1(\mathbf{x}^{*b}) - m_1^*]^2}{B-1}} = 13.32$$

$$\text{avec } m_1^* = \sum_b m_1(\mathbf{x}^{*b})/B = 55.73$$

Exemples d'estimation d'erreurs standard

durée de survie

groupe 1 (placébo)	groupe 2 (traitement)
$n_1 = 9$ mesures	$n_2 = 7$ mesures
52, 10, 40, 104, 50, 27, 146, 31, 46	94, 38, 23, 197, 99, 16, 141
moyenne $m_1 = 56.22$	moyenne $m_2 = 86.86$
médiane $\theta_1 = 46$	médiane $\theta_2 = 94$

erreur standard sur m_1 :
- ➥ classique : $se_1 = 14.14$
- ➥ bootstrap : $se_1^* = 13.32$

erreur standard sur m_1 :
- ➥ classique : $se_2 = 25.24$
- ➥ bootstrap : $se_2^* = 23.81$

erreur standard sur θ_1 :
- ➥ classique : ?
- ➥ bootstrap : $se_1^* = 11.54$

erreur standard sur θ_2 :
- ➥ classique : ?
- ➥ bootstrap : $se_2^* = 36.35$

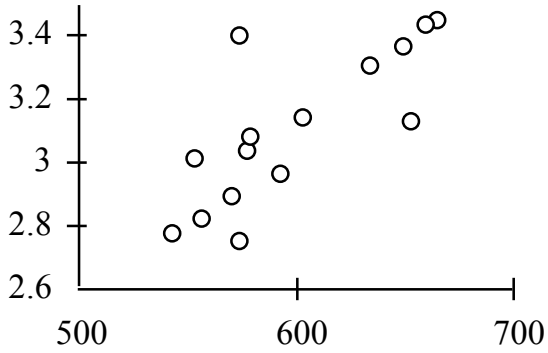
erreur standard sur n'importe quelle statistique
- ➥ classique : ?
- ➥ bootstrap : TOUJOURS UNE SOLUTION

au prix d'un peu de calcul ...

Erreur standard d'un coefficient de corrélation (1)

performances à des tests de contrôle de connaissance

test national précédent la scolarisation	note moyenne dans l'année qui suit
576	3.39
635	3.30
558	2.81
578	3.03
666	3.44
580	3.07
555	3.00
661	3.43
651	3.36
605	3.13
653	3.12
575	2.74
545	2.76
572	2.88
594	2.96



r=0.776

fiabilité de cette valeur ?

➥ bootstrap

Erreur standard d'un coefficient de corrélation (2)

échantillon observé

$$x=\begin{pmatrix} 576 & 635 & 558 & 578 & 666 & 580 & 555 & 661 & 651 & 605 & 653 & 575 & 545 & 572 & 594 \\ 3.39 & 3.30 & 2.81 & 3.03 & 3.44 & 3.07 & 3.00 & 3.43 & 3.36 & 3.13 & 3.12 & 2.74 & 2.76 & 2.88 & 2.96 \end{pmatrix}$$

statistique d'intérêt : corrélation r=0.776

B échantillons bootstrap

réplications bootstrap de la corrélation r

calcul de la corrélation r

$$x^{*1}=\begin{pmatrix} 661 & 558 & 666 & 651 & \dots & 594 \\ 3.43 & 2.81 & 3.44 & 3.36 & \dots & 2.96 \end{pmatrix} \rightarrow 0.927$$

$$x^{*b}=\begin{pmatrix} 651 & 575 & 605 & 575 & \dots & 575 \\ 3.36 & 2.74 & 3.13 & 2.74 & \dots & 2.74 \end{pmatrix} \rightarrow 0.900$$

…

$$x^{*B}=\begin{pmatrix} 572 & 572 & 545 & 653 & \dots & 575 \\ 2.88 & 2.88 & 2.76 & 3.12 & \dots & 2.74 \end{pmatrix} \rightarrow 0.793$$

$$SE(r) = \sqrt{\frac{\sum_b [r(x^{*b}) - r^*]^2}{B-1}} = 0.775$$

avec $r^* = \sum_b r(x^{*b})/B = 0.134$

Erreurs standard en ACP (1)

élève	notes par matière				
	math	phys	litt	angl	mus
1	17	14	18	14	12
2	09	13	15	16	18
.					
.					
.					
i	x_{i1}	x_{i2}	x_{ij}		x_{i5}
.					
.					
.					
N	19	15	09	12	06

- Matrice 5x5 de covariance empirique **G** :

$$G_{jk} = \frac{1}{N} \sum_i [x_{ij} - moy_i(x_{ij})] [x_{ik} - moy_i(x_{ik})] \quad j,k=1...5$$

- Calcul des valeurs propres et vecteurs propres de **G** :

$\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5$ et $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \mathbf{v}_4, \mathbf{v}_5$

fiabilité du pourcentage d'inertie $\lambda_l / \sum_k \lambda_k$?

fiabilité des $\mathbf{v}_k$?

➥ bootstrap

Erreurs standard en ACP (2)

échantillon observé

$\mathbf{X} =$

élève	notes par matière				
	math	phys	litt	angl	mus
1	17	14	18	14	12
2	09	13	15	16	18
.					
.					
.					
i	x_{i1}	x_{i2}	x_{ij}		x_{i5}
.					
.					
.					
N	19	15	09	12	06

B échantillons bootstrap

statistiques d'intérêt :
%age d'inertie PI = $\lambda_1/\sum_k \lambda_k$
vecteurs propres $\mathbf{v}_k$

$\mathbf{X^{*1}} =$

élève	notes par matière				
	math	phys	litt	angl	mus
1	08	11	19	17	15
2	09	13	15	16	18
.					
.					
.					
i	x_{i1}	x_{i2}	x_{ij}		x_{i5}
.					
.					
.					
N	17	14	18	14	12

calcul de $\mathbf{G^{*b}}$, valeurs propres et vecteurs propres de G

réplications bootstrap

$\lambda_1^{*1}/\sum_k \lambda_k^{*1}$
$\mathbf{v}_1^{*1}, \mathbf{v}_2^{*1}, \mathbf{v}_3^{*1}, \mathbf{v}_4^{*1}, \mathbf{v}_5^{*1}$

...

...

$\mathbf{X^{*B}} =$

élève	notes par matière				
	math	phys	litt	angl	mus
1	09	13	15	16	18
2					
.					
.					
.					
i	x_{i1}	x_{i2}	x_{ij}		x_{i5}
.					
.					
.					
N	08	11	19	17	15

$\lambda_1^{*B}/\sum_k \lambda_k^{*B}$
$\mathbf{v}_1^{*B}, \mathbf{v}_2^{*B}, \mathbf{v}_3^{*B}, \mathbf{v}_4^{*B}, \mathbf{v}_5^{*B}$

$$SE(PI) = \sqrt{\frac{\sum_b [PI(\mathbf{X^{*b}}) - PI^*]^2}{B-1}} \quad \text{avec } PI^* = \sum_b PI(\mathbf{x^{*b}})/B$$

$$SE(\mathbf{v}_k) = \sqrt{\frac{\sum_b [\mathbf{v}_k(\mathbf{X^{*b}}) - \mathbf{v}_k^*]^2}{B-1}} \quad \text{avec } \mathbf{v}_k^* = \sum_b \mathbf{v}_k(\mathbf{X^{*b}})/B$$

Erreur standard dans l'ajustement de courbes (1)

Diminution du taux de cholestérol (y) en fonction du pourcentage de la dose prescrite effectivement absorbée (x)

x_i(%)	0	2	7	8	16	33	43	...	100
y_i	11.5	5.75	-10.5	36.25	29.75	27.75	33.25		86.75

- Modèle

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2$$

- Ajustement des moindres carrés

➥ $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$

- Diminution prédite par le modèle

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_2 x_i^2$$

fiabilité des valeurs prédites, i.e., erreur standard autour d'une valeur prédite pour le modèle considéré ?
e.g., erreur standard autour de $y_{60\%}$?
➥ bootstrap

Erreur standard dans l'ajustement de courbes (2)

1ère approche

échantillon observé

x_i(%)	0	2	7	8	16	33	43	...	100
y_i	11.5	5.75	-10.5	36.25	29.75	27.75	33.25		86.75

statistiques d'intérêt : valeurs prédites $\hat{y}_i$

B échantillons bootstrap

calcul de $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$

réplications bootstrap de $\hat{y}_i$

x^{*1}	0	54	43	2	...	16
y^{*1}	11.5	47.25	33.25	5.75	...	29.75

→ y_i^{*1}

x^{*b}	33	95	7	43	...	72
y^{*b}	27.75	77.00	-10.5	33.25	...	63.00

→ y_i^{*b}

...

x^{*B}	100	72	43	28	...	7
y^{*B}	86.75	63.00	33.25	23.5	...	-10.5

→ y_i^{*B}

$$SE(\hat{y}_i) = \sqrt{\frac{\sum_b [y_i^{*b} - y_i^*]^2}{B-1}}$$

avec $y_i^* = \sum_b y_i^{*b} / B$

Erreur standard dans l'ajustement de courbes (3)

2ème approche

échantillon observé

x_i(%)	0	2	7	8	16	33	43	...	100
y_i	11.5	5.75	-10.5	36.25	29.75	27.75	33.25		86.75

ajustement du modèle : $y_i = \beta_0+\beta_1 x_i+\beta_2 x_i^2$

statistiques d'intérêt : valeurs prédites $\hat{y}_i$

$\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$

1 échantillon de résidus : $\hat{\varepsilon}_i = y_i-\hat{\beta}_0+\hat{\beta}_1 x_i+\hat{\beta}_2 x_i^2$

$\hat{\varepsilon}_i$ = 1.2 2.4 -1.3 ... -0.8

B échantillons bootstrap de résidus

modèle : $y_i^{*b}=\hat{\beta}_0+\hat{\beta}_1 x_i+\hat{\beta}_2 x_i^2+\varepsilon_i^{*b}$

ε_i^{*1}	2.4	-1.3	0.7	...	0.6	y_i^{*1}
ε_i^{*b}	-1.3	-0.8	1.6	...	1.2	y_i^{*b}
ε_i^{*B}	2.4	1.2	0.5	...	-0.1	y_i^{*B}

B réplications bootstrap de $\hat{y}_i$

erreur standard de $\hat{y}_i$

Ajustement de courbes : résumé

2 possibilités :

• Bootstrap des paires (x_i, y_i)

➥ pas de modèle nécessaire

➥ suppose que les paires sont des réalisations aléatoires de la population

• Bootstrap des résidus

➥ sensible au modèle

Si modèle incertain,
adopter plutôt le bootstrap des paires

Nombre B de réplications bootstrap nécessaire

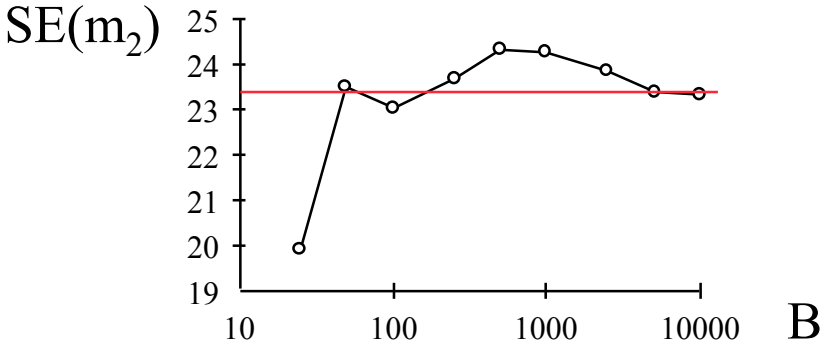
REGLES EMPIRIQUES

• Même un petit nombre de réplications fournit déjà des informations très utiles. B=50 est souvent suffisant pour une estimation fiable de l'erreur standard

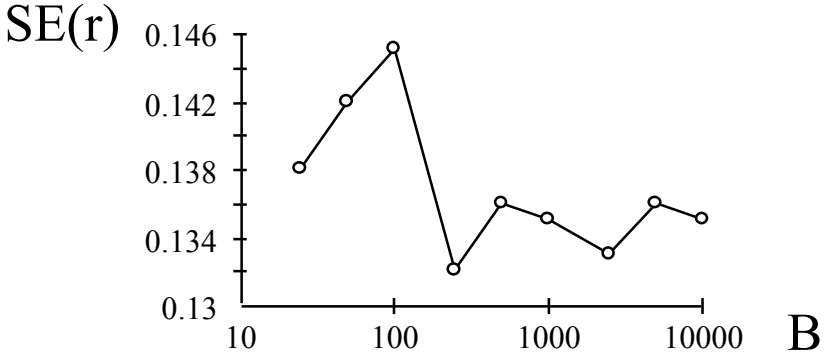
• Il est rare que plus de 200 réplications soient nécessaires pour estimer les erreurs standard

Exemples :

erreur standard de la moyenne m_2



erreur standard du coefficient de corrélation r



Type de données : structurées vs non structurées

• Données non structurées

➥ les valeurs de l'échantillon observé sont indépendantes

➥ une modification de l'ordre des valeurs ne modifie pas l'échantillon

➥ exemples :

durée de survie des animaux
notes des étudiants aux tests
notes des étudiants dans les différentes disciplines

• Données structurées

➥ les valeurs de l'échantillon observé ne sont pas indépendantes

➥ l'ordre des valeurs dans l'échantillon est important

➥ exemples :

série temporelle ou chronologique
spectre en énergie
image

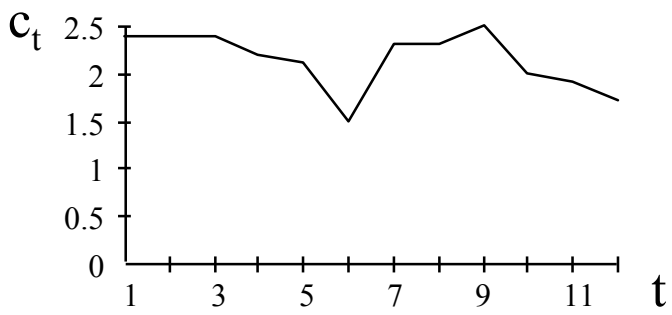
ATTENTION

Dans le cas de données structurées, la procédure de calcul d'échantillons bootstrap ne doit pas détruire la structure !

Bootstrap d'une série temporelle : problème

Evolution de la concentration d'une hormone au cours du temps

t	1	2	3	4	5	6	7	8	9	10	11	12
c_t	2.4	2.4	2.4	2.2	2.1	1.5	2.3	2.3	2.5	2.0	1.9	1.7



- Modèle

  centrage des mesures : $y_t = c_t - \text{moy}(c_t)$
  modèle AR1 : $y_t = \square\, y_{t-1} + \square_t$

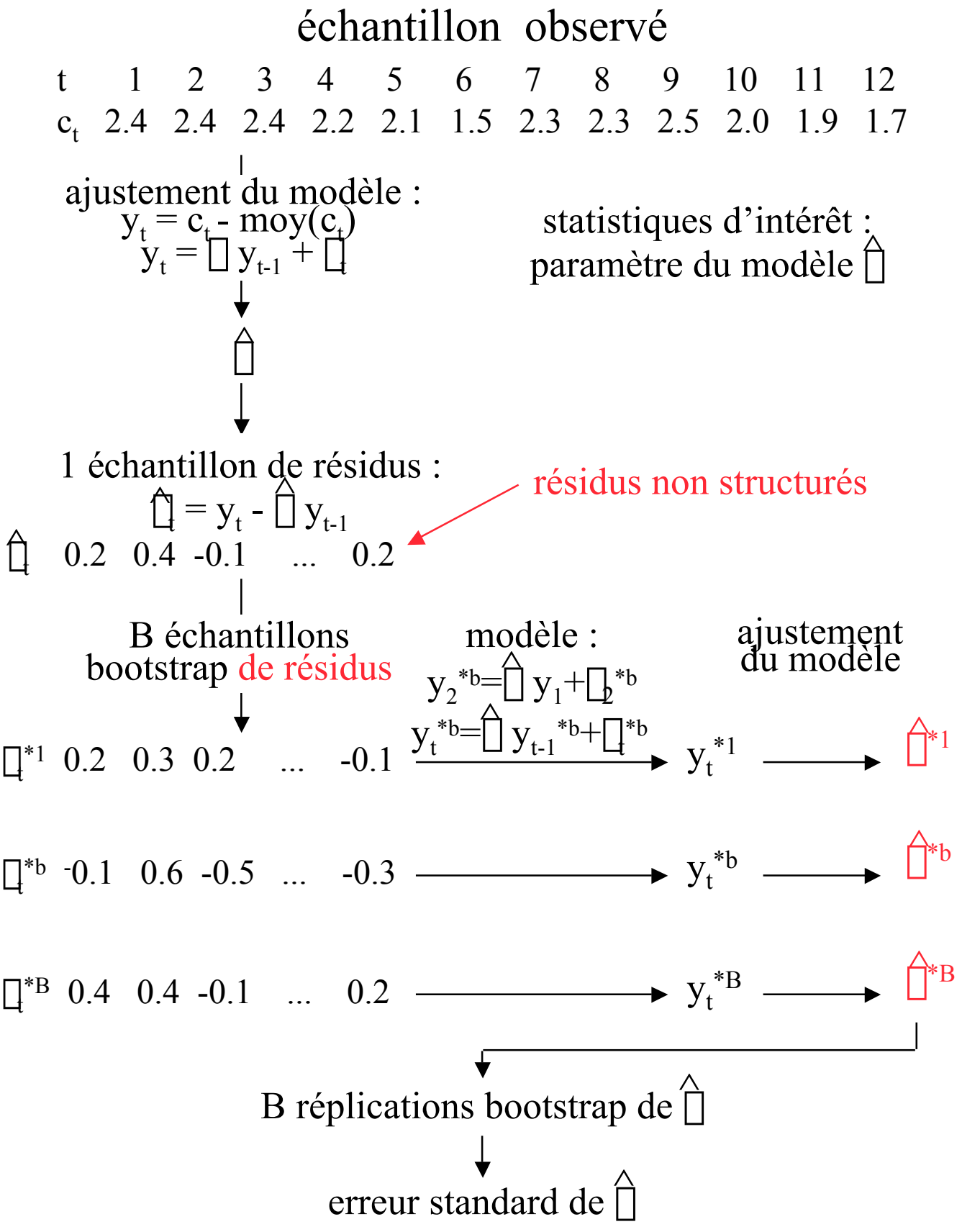
- Ajustement des moindres carrés

  ➥ $\hat{\square}$

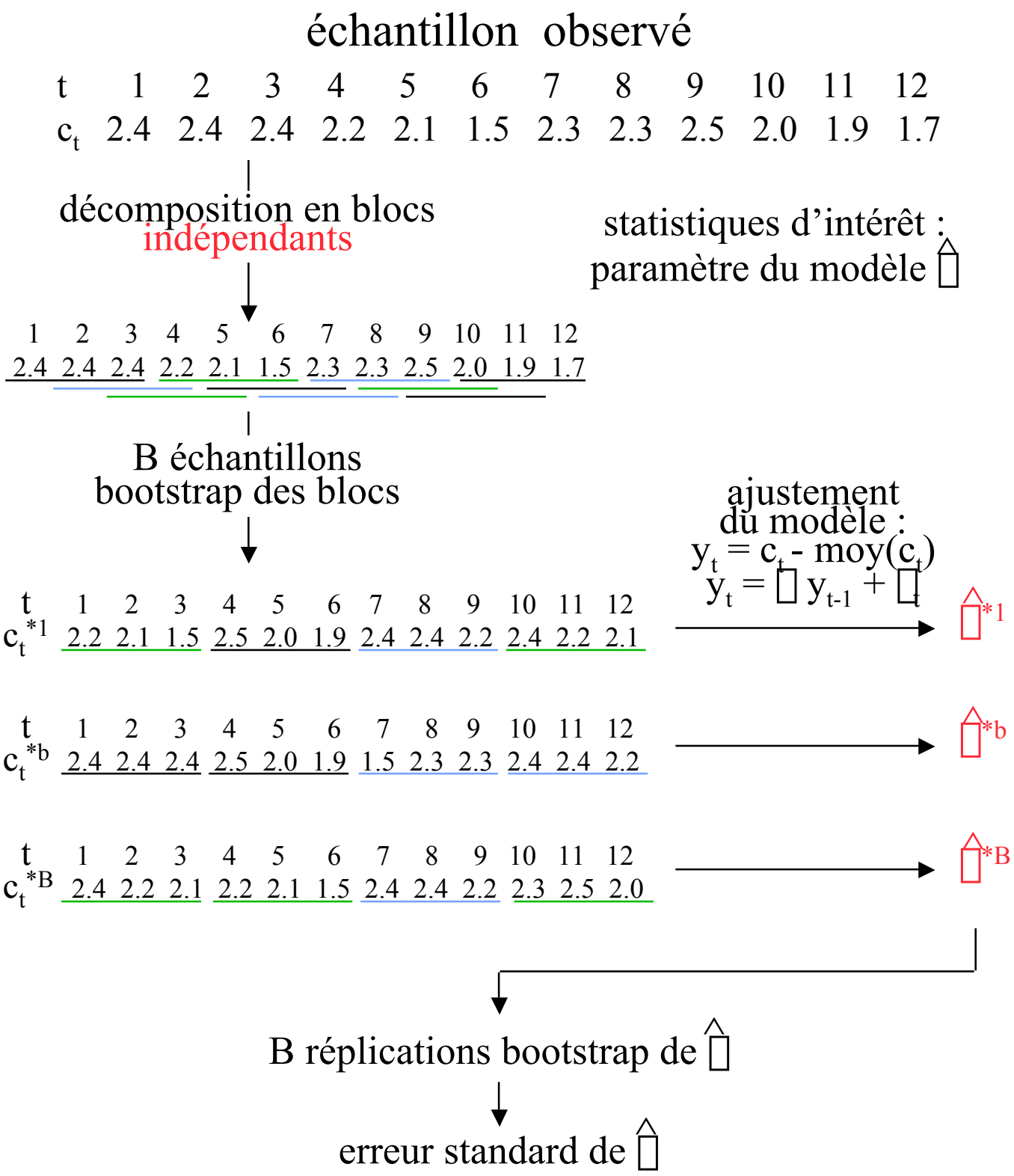
Fiabilité de $\hat{\square}$?

➥ bootstrap

Bootstrap d'une série temporelle : 1ère approche



Bootstrap d'une série temporelle : 2ème approche



Bootstrap d'une série temporelle : résumé

2 possibilités :

• Modèle et bootstrap des résidus

➥ modèle tel que les résidus soient non structurés

➥ bootstrap des résidus

➥ reconstitution de données structurées bootstrap à partir du modèle et des réplications bootstrap des résidus

➥ estimation de la statistique d'intérêt sur chaque série temporelle bootstrap reconstituée

• Bootstrap par blocs

➥ décomposition de la série en blocs indépendants

➥ reconstitution de séries bootstrap en joignant les blocs tirés aléatoirement avec remise

➥ estimation de la statistique d'intérêt sur chaque série temporelle bootstrap reconstituée

➥ moins dépendant d'un modèle, mais problème du choix de la longueur des blocs

Bootstrap pour l'estimation du biais : 1ère approche

biais = valeur estimée - valeur vraie

1 échantillon observé $\mathbf{x} = (x_1, x_2, \ldots, x_N)$ et 1 statistique d'intérêt $s(\mathbf{x})$: moyenne, médiane,...

↓

B échantillons bootstrap

↓

		réplications bootstrap de s
	calcul de la statistique d'intérêt	
$\mathbf{x}^{*1} = (x_1^*, x_2^*, \ldots, x_N^*)$	→	$s(\mathbf{x}^{*1})$
$\mathbf{x}^{*b} = (x_1^*, x_2^*, \ldots, x_N^*)$	→	$s(\mathbf{x}^{*b})$
…		…
$\mathbf{x}^{*B} = (x_1^*, x_2^*, \ldots, x_N^*)$	→	$s(\mathbf{x}^{*B})$

➥ estimée bootstrap du biais

$$\text{biais} = s^* - s(\mathbf{x})$$

$$\text{avec } s^* = \sum_b s(\mathbf{x}^{*b}) / B$$

Vecteur de rééchantillonnage

1 échantillon observé

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

|

1 échantillon bootstrap $\mathbf{x}^{*b}$ $\Leftrightarrow$ 1 vecteur de rééchantillonnage $\mathbf{P}^{*b}$

↓

$\mathbf{x}^{*b} = (x_1^* , x_2^* , \ldots , x_N^*)$

$P_j^{*b} = \#(x_j^* = x_j)/N \quad j=1,\ldots,N$

= nb d'occurrences de x_j dans l'échantillon bootstrap

Exemple :

$\mathbf{x} = (x_1 , x_2 , x_3 , x_4 , x_5 , x_6 , x_7 , x_8)$

$\mathbf{x}^{*1} = (x_3 , x_2 , x_7 , x_7 , x_4 , x_3 , x_3 , x_7)$

$\mathbf{P}^{*1} = (0, 1/7, 3/7, 1/7, 0, 0, 3/7, 0)$

1 réplication bootstrap de la statistique $s(\mathbf{x}^{*b})$

$\Leftrightarrow$ 1 fonction $S(\mathbf{P}^{*b})$ du vecteur de rééchantillonnage $\mathbf{P}^{*b}$

Exemple :

$s(\mathbf{x}^{*b})$ = moyenne de l'échantillon = $\sum_J \mathbf{x}_j^{*b} /N$

$\Leftrightarrow \quad S(\mathbf{P}^{*b}) = \sum_J \mathbf{x}_j \mathbf{P}_j^{*b}$

Bootstrap pour l'estimation du biais : 2ème approche

1 échantillon observé $\mathbf{x} = (x_1, x_2, \ldots, x_N)$ et 1 statistique d'intérêt $s(\mathbf{x})$: moyenne, médiane,...

↓

B échantillons bootstrap

↓

calcul du vecteur de rééchantillonnage et calcul de la statistique d'intérêt

$\mathbf{x}^{*1} = (x_1^*, x_2^*, \ldots, x_N^*) \rightarrow P^{*1}, s(x^{*1})$

$\mathbf{x}^{*b} = (x_1^*, x_2^*, \ldots, x_N^*) \rightarrow P^{*b}, s(x^{*b})$

…

$\mathbf{x}^{*B} = (x_1^*, x_2^*, \ldots, x_N^*) \rightarrow P^{*B}, s(x^{*B})$

➥ moyenne du vecteur d'échantillonnage

$$P^* = \sum_b P^{*b} / B$$

➥ moyenne des réalisations bootstrap de la statistique

$$s^* = \sum_b s(\mathbf{x}^{*b}) / B$$

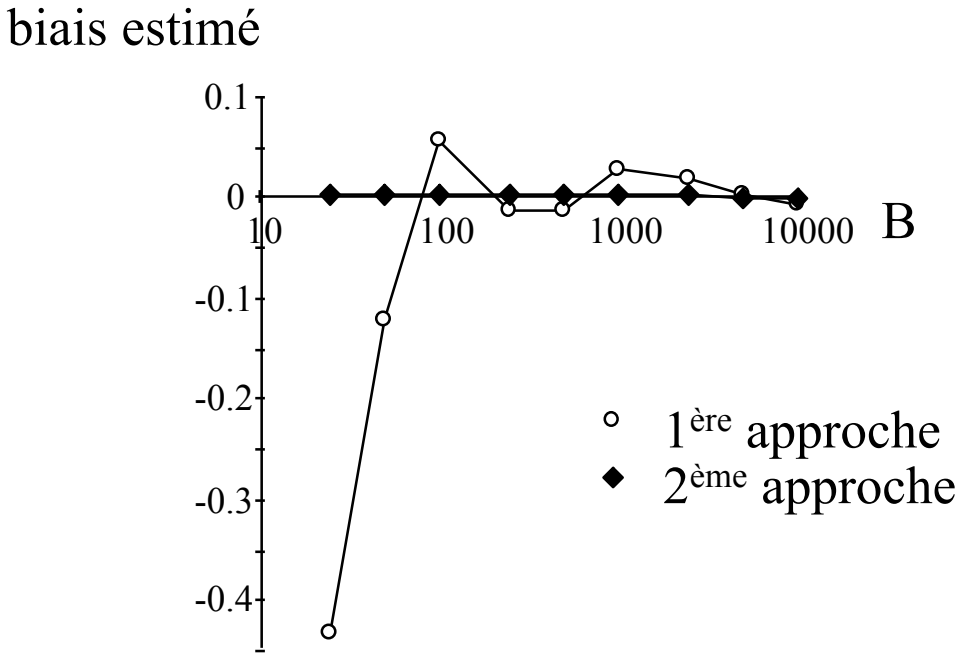
➥ estimée bootstrap du biais

$$\text{biais} = s^* - S(\mathbf{P}^*)$$

Bootstrap pour l'estimation du biais : exemple

échantillon observé

x = (26, 27, 29, 36, 35, 33, 35, 24, 31, 34, 42, 28, 35, 35, 27)
statistique d'intérêt : moyenne m = 31.80
valeur vraie = 30



➥ convergence des deux approches
➥ convergence beaucoup plus rapide de la 2ème approche
➥ à la convergence, possible écart par rapport à la valeur vraie, inhérent à l'estimation à partir d'un échantillon fini

Correction du biais par l'approche bootstrap

biais = valeur estimée - valeur vraie

s_{corr} = s(**x**) - biais estimé

= 2s(**x**) - s^* ($1^{ère}$ approche)

= s(**x**) - s^* + S($\mathbf{P}^*$) ($2^{ère}$ approche)

ATTENTION

➥ l'estimation corrigée du biais n'est pas s^*

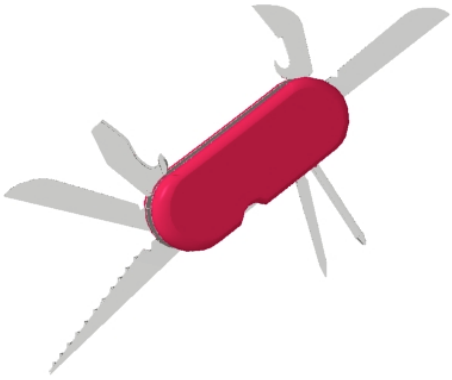
➥ la correction de biais peut être dangereuse en pratique car s_{corr} peut avoir une grande erreur standard

RECOMMANDATIONS

➥ si biais faible par rapport à l'erreur standard, mieux vaut utiliser s(**x**) plutôt que s_{corr}

➥ si biais grand par rapport à l'erreur standard, s(**x**) n'est probablement pas une bonne approximation de la statistique d'intérêt pour la population

Bootstrap ou Jackknife ?



Définition d'un échantillon jackknife

1 échantillon observé de N valeurs

$\mathbf{x} = (x_1, x_2, x_3, \ldots x_i \ldots x_N)$

$\mathbf{x} = (50, 53, 58, 80, 75, 69, 77, 44, 63, 73)$

|

échantillon jackknife $\mathbf{x}_i$:
échantillon original sans l'observation i

↓

$\mathbf{x}_i = (x_1, x_2, x_3, \ldots x_{i-1}, x_{i+1}, \ldots x_N)$

$\mathbf{x}_3 = (50, 53, 80, 75, 69, 77, 44, 63, 73)$

- à partir d'un échantillon observé contenant N valeurs
  ➥ N échantillons jackknife seulement

Estimation jackknife de l'erreur standard et du biais

- Statistique d'intérêt s

- Estimation jackknife de l'erreur standard de s

$$SE_{jackknife}(s) = \sqrt{\frac{N-1}{N} \sum_{1} [\, s(\mathbf{x}_i\,) - s\,]^2}$$

$$\text{avec} \quad s\, = \sum_{1} s(\mathbf{x}_i\,)/N$$

à comparer à :

$$SE_{bootstrap}(s) = \sqrt{\frac{\sum_{b} [s(\mathbf{x}^{*b}) - s^*]^2}{B-1}}$$

➥ facteur d'inflation (N-1)/N requis car les échantillons jackknife sont moins dissemblables de l'échantillon initial que les échantillons bootstrap

- Estimation jackknife du biais

$$biais_{jackknife}(s) = (N-1)\,[s\, - s(\mathbf{x})\,]$$

Jackknife versus bootstrap

• Travaux jackknife préalables aux travaux bootstrap

• Jackknife = approximation du bootstrap

- statistique linéaire s(**x**) = constante + $\sum$ fonction(x_i)

➥ pas de perte d'information par l'approche jackknife

- statistique non linéaire s(**x**)

➥ perte d'informations par l'approche jackknife

➥ jackknife = approximation linéaire du bootstrap

• Jackknife = moins efficace que le bootstrap en général

➥ écart entre estimées bootstrap et jackknife fonction de l'écart de la statistique d'intérêt à la linéarité

• Echec du jackknife si la statistique d'intérêt n'est pas une fonction différentiable de **x** (par exemple, médiane)

RECOMMANDATION :

➥ préférer l'approche bootstrap !

Bootstrap et estimation d'intervalles de confiance

$$\text{Prob}\,(s \in [s_1 ; s_2]) = 1 - 2\alpha$$

• Plusieurs approches possibles :

- construction de tables bootstrap

➥ non recommandée pour les problèmes non paramétriques

- utilisation des percentiles bootstrap

➥ juste au premier ordre :

$\text{prob}(s<s_1) = \alpha + c_1 / \sqrt{N}$ et $\text{prob}(s>s_2) = \alpha\ c_2 / \sqrt{N}$

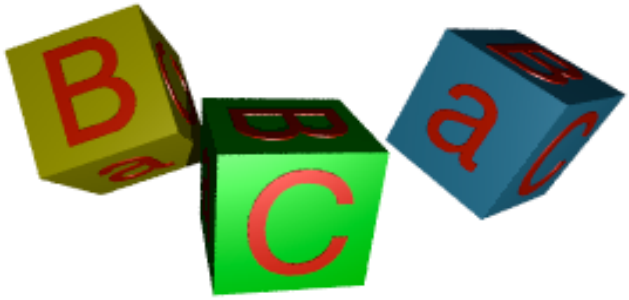
- méthode BC_a : Bias-Corrected and accelerated

➥ juste au second ordre :

$\text{prob}(s<s_1) = \alpha + c_1 / N$ et $\text{prob}(s>s_2) = \alpha\ c_2 / N$

➥ plus qu'un avantage théorique

➥ méthode recommandée



Méthode des percentiles bootstrap

1 échantillon observé $\mathbf{x} = (x_1, x_2, \ldots, x_N)$ et 1 statistique d'intérêt $s(\mathbf{x})$: moyenne, médiane,...

↓

B échantillons bootstrap

↓

calcul de la statistique d'intérêt

B réplications bootstrap de s

$\mathbf{x}^{*1} = (x_1^*, x_2^*, \ldots, x_N^*) \longrightarrow s(\mathbf{x}^{*1})$

$\mathbf{x}^{*b} = (x_1^*, x_2^*, \ldots, x_N^*) \longrightarrow s(\mathbf{x}^{*b})$

…

$\mathbf{x}^{*B} = (x_1^*, x_2^*, \ldots, x_N^*) \longrightarrow s(\mathbf{x}^{*B})$

- Classement des B valeurs de $s(\mathbf{x}^{*b})$ par ordre croissant
- Intervalle de confiance $[s_1 ; s_2]$ couvrant $1-2\alpha$, i.e.,

$$\text{Prob}(s \in [s_1; s_2]) = 1-2\alpha$$

intervalle contenant $100 * (1-2\alpha)\%$ des valeurs

avec : $s_1 = 100.\alpha^{\text{ième}}$ percentile des $s(\mathbf{x}^{*b})$ calculés, i.e., $B.\alpha^{\text{ième}}$ valeur de la liste classée par ordre croissant

$s_2 = 100.(1-\alpha)^{\text{ième}}$ percentile des $s(\mathbf{x}^{*b})$ calculés, i.e., $B.(1-\alpha)^{\text{ième}}$ valeur de la liste classée par ordre croissant

- Exemple :

$B = 2000$ et $\alpha = 5\%$

$s_1 = 100^{\text{ème}}$ valeur de la liste classée

$s_2 = 1900^{\text{ème}}$ valeur de la liste classée

Méthode BC_a

$$\text{Prob}\,(s \in [s_1 ; s_2]) = 1 - 2\alpha$$

• Bornes s_1 et s_2 également exprimées à partir des percentiles de la distribution bootstrap

• Bornes s_1 et s_2 différentes de celles de la méthode des percentiles :

$s_1 = 100.\alpha_1^{\text{ième}}$ percentile des $s(\mathbf{x}^{*b})$ calculés, i.e., $B.\alpha_1^{\text{ième}}$ valeur de la liste classée par ordre croissant

$s_2 = 100.\alpha_2^{\text{ième}}$ percentile des $s(\mathbf{x}^{*b})$ calculés, i.e., $B.\alpha_2^{\text{ième}}$ valeur de la liste classée par ordre croissant

avec :

$$\alpha_1 = \Phi\left(z_0 + \frac{z_0 + z^{(\alpha)}}{1 - a\,(z_0 + z^{(\alpha)})}\right)$$

$$\alpha_2 = \Phi\left(z_0 + \frac{z_0 + z^{(1-\alpha)}}{1 - a\,(z_0 + z^{(1-\alpha)})}\right)$$

où : Φ est la fonction de distribution cumulée de la loi normale centrée réduite, e.g., $\Phi(1.645) = 0.95$

$z^{(\alpha)}$ est le $100.\alpha^{\text{ième}}$ percentile de la loi normale centrée réduite, e.g., $z^{(0.95)} = 1.645$

$z_0 = \Phi^{-1}[$ (nb de valeurs $s(\mathbf{x}^{*b}) < s(\mathbf{x}))/B]$

Φ^{-1} est l'inverse de la fonction de distribution cumulée de la loi normale centrée réduite, e.g., $\Phi^{-1}(0.95) = 1.645$

$$a_0 = \frac{\sum_1 [s_{\cdot} - s(\mathbf{x}_i)]^3}{6\,\{\sum_1 [s_{\cdot} - s(\mathbf{x}_i)]^2\}^{3/2}}$$

Nombre d'échantillons bootstrap nécessaires

ATTENTION

➥ plus de 1000 échantillons bootstrap sont nécessaires pour une estimation robuste des intervalles de confiance

Bootstrap et tests d'hypothèse

- Les 2 échantillons observés émanent t-il de la même distribution de probabilité ?

- Les moyennes des deux populations sous-jacentes à deux échantillons observés sont-elles identiques ?

- La moyenne des observations est-elle significativement différente d'une valeur théorique ?

➥ l'approche bootstrap peut répondre !

Notion de niveau de signification atteint (ASL)

Niveau de signification atteint
= Achieved Significance Level ASL

Probabilité d'observer une valeur de test au moins aussi grande que la valeur observée quand l'hypothèse H0 est vraie

$$ASL = Prob_{H0}(t^* \geq t_{obs})$$

• Plus ASL est faible, plus il y a d'évidence pour rejeter H0

• Si ASL < □, rejeter H0

• La valeur t_{obs} est fixe et correspond à la valeur de test calculée à partir de ou des échantillons effectivement observés

• La valeur t* correspond à la valeur de test sous l'hypothèse H0, estimé par le bootstrap

Tests d'hypothèse : principe général

- Nécessité de définir 2 quantités :
  - ➥ une statistique de test t
  - ➥ la distribution des données F_0 sous l'hypothèse H0

- Générer B échantillons bootstrap de $t(x^*)$ à partir de la distribution F_0

- Calculer le niveau de signification atteint par
  ASL = (nb de valeurs $t(\mathbf{x}^{*b}) \geq t_{obs}$)/B

- Si ASL < □, rejeter H0

Tests d'hypothèse : exemple 1

2 échantillons observés

$\mathbf{y} = (y_1, y_2, \ldots, y_N)$, $\text{moy}(\mathbf{y}) = \sum_i y_i / N$

$\mathbf{z} = (z_1, z_2, \ldots, z_M)$, $\text{moy}(\mathbf{z}) = \sum_i z_i / M$

Les 2 échantillons $\mathbf{y}$ et $\mathbf{z}$ observés émanent t-il de la même distribution de probabilité F_0 ?

H0 : y et z sont des échantillons issus d'une même population de distribution F_0

- Former $\mathbf{x} = (\mathbf{y}, \mathbf{z})$
- Tirer B échantillons bootstrap de taille N+M à partir de $\mathbf{x}$.

Pour chaque échantillon, les N premières observations sont notées $\mathbf{y}^{*b}$ et les M suivantes $\mathbf{z}^{*b}$.

- Pour chaque échantillon bootstrap, calculer :

  $t(\mathbf{x}^{*b}) = \text{moy}(\mathbf{y}^{*b}) - \text{moy}(\mathbf{z}^{*b})$

  avec $\text{moy}(\mathbf{y}^{*b}) = \sum_i y_i^{*b} / N$ et $\text{moy}(\mathbf{z}^{*b}) = \sum_i z_i^{*b} / M$

- Calculer le niveau de signification atteint par

  $\text{ASL} = (\text{nb de valeurs } t(\mathbf{x}^{*b}) \geq t_{obs})/B$

  où $t_{obs} = \text{moy}(\mathbf{y}) - \text{moy}(\mathbf{z})$

- Si ASL < α, rejeter H0

Rq : une autre statistique de test peut être utilisée à la place de $t(\mathbf{x}^{*b}) = \text{moy}(\mathbf{y}^{*b}) - \text{moy}(\mathbf{z}^{*b})$, par exemple une statistique de Student

Tests d'hypothèse : exemple 2

2 échantillons observés

$\mathbf{y} = (y_1, y_2, \ldots, y_N)$, $\text{moy}(\mathbf{y}) = \sum_i y_i / N$

$\mathbf{z} = (z_1, z_2, \ldots, z_M)$, $\text{moy}(\mathbf{z}) = \sum_i z_i / M$

Les 2 échantillons **y** et **z** observés émanent t-il de populations présentant la même moyenne ?

H0 : moy(**y**) = moy(**z**)

- Former $\mathbf{x} = (\mathbf{y}, \mathbf{z})$ et calculer $\text{moy}(\mathbf{x}) = \sum_i y_i / N$
- Calculer $y'_i = y_i - \text{moy}(\mathbf{y}) + \text{moy}(\mathbf{x})$
  et $z'_i = z_i - \text{moy}(\mathbf{z}) + \text{moy}(\mathbf{x})$
- Tirer B échantillons bootstrap $\mathbf{y}^{*b}$ de taille N à partir de **y'**, B échantillons bootstrap $\mathbf{z}^{*b}$ de taille M à partir de **z'**. En déduire B vecteurs $\mathbf{x}^{*b} = (\mathbf{y}^{*b}, \mathbf{z}^{*b})$
- Pour chaque échantillon bootstrap, calculer :

$$t(\mathbf{x}^{*b}) = \frac{\text{moy}(\mathbf{y}^{*b}) - \text{moy}(\mathbf{z}^{*b})}{\sqrt{\sigma_y^{2*b}/N + \sigma_z^{2*b}/M}} \quad \text{avec}$$

$$\text{moy}(\mathbf{y}^{*b}) = \sum_i y_i^{*b} / N \quad \text{et} \quad \text{moy}(\mathbf{z}^{*b}) = \sum_i z_i^{*b} / M$$

$$\sigma_y^{2*b} = \sum_i (y_i^{*b} - \text{moy}(\mathbf{y}^{*b}))^2/(N-1)$$

$$\sigma_z^{2*b} = \sum_i (z_i^{*b} - \text{moy}(\mathbf{z}^{*b}))^2/(M-1)$$

- Calculer le niveau de signification atteint par

$$\text{ASL} = (\text{nb de valeurs } t(\mathbf{x}^{*b}) \geq t_{obs})/B$$

$$\text{où } t_{obs} = \frac{\text{moy}(\mathbf{y}) - \text{moy}(\mathbf{z})}{\sqrt{\sigma_y^2/N + \sigma_z^2/M}}$$

Tests d'hypothèse : exemple 3

1 échantillon observé

$\mathbf{x} = (x_1 , x_2 , \ldots , x_N)$, $\text{moy}(\mathbf{x}) = \sum_i x_i /N$

La moyenne de l'échantillon observé vaut-elle μ ?

H0 : $\text{moy}(\mathbf{x}) = \mu$

- Tirer B échantillons bootstrap $\mathbf{x}^{*b}$ de taille N à partir de $\mathbf{x}$

- Pour chaque échantillon bootstrap, calculer :

$$t(\mathbf{x}^{*b}) = \frac{\text{moy}(\mathbf{x}^{*b}) - \text{moy}(\mathbf{x})}{\sqrt{\sigma^{2*b}/N}}$$

avec $\quad \text{moy}(\mathbf{x}^{*b}) = \sum_i x_i^{*b} /N$

$\sigma^{2*b} = \sum_i (x_i^{*b} - \text{moy}(\mathbf{x}^{*b}))^2/(N-1)$

- Calculer le niveau de signification atteint par

ASL = (nb de valeurs $t(\mathbf{x}^{*b}) \geq t_{obs}$)/B

où $t_{obs} = \dfrac{\text{moy}(\mathbf{x}) - \mu}{\sqrt{\sigma^2/N}}$

- Si ASL < α, rejeter H0

Bootstrap paramétrique

1 échantillon observé de N valeurs

x = (50 ; 53 ; 58 ; 80 ; 75 ; 69 ; 77 ; 44 ; 63 ; 73)

non paramétrique

↓

1 échantillon bootstrap :
1 tirage aléatoire de N valeurs parmi l'échantillon original, avec remise

paramétrique

↓

estimation de la loi de la population

↓

1 échantillon bootstrap :
1 tirage aléatoire de N valeurs à partir de la loi de la population

- Bootstrap non paramétrique
  - ➥ aucune hypothèse de loi de la population sous-jacente nécessaire

- Bootstrap paramétrique
  - ➥ moins biaisé que les expressions analytiques
  - ➥ fournit des solutions aux problèmes pour lesquels il n'existe pas de formule analytique

Bilan

• Bootstrap = méthode d'inférence statistique adaptée au contexte non paramétrique

• 1 seul échantillon d'observations nécessaire

• Permet d'estimer la distribution sous-jacente à une population

• Permet d'associer des erreurs standard à virtuellement n'importe quelle statistique :

➥ moyenne, médiane

➥ coefficient de corrélation

➥ paramètres issus d'une modélisation des données

➥ analyse multidimensionnelle (ACP)

• Permet d'étudier le biais associé à une statistique calculée à partir d'un seul échantillon

• Permet de calculer des intervalles de confiance et de réaliser des tests d'hypothèse

• Estimateurs bootstrap = estimateurs non biaisés

Sujets plus avancés relatifs au bootstrap

• Estimation de la puissance d'un test à partir du bootstrap

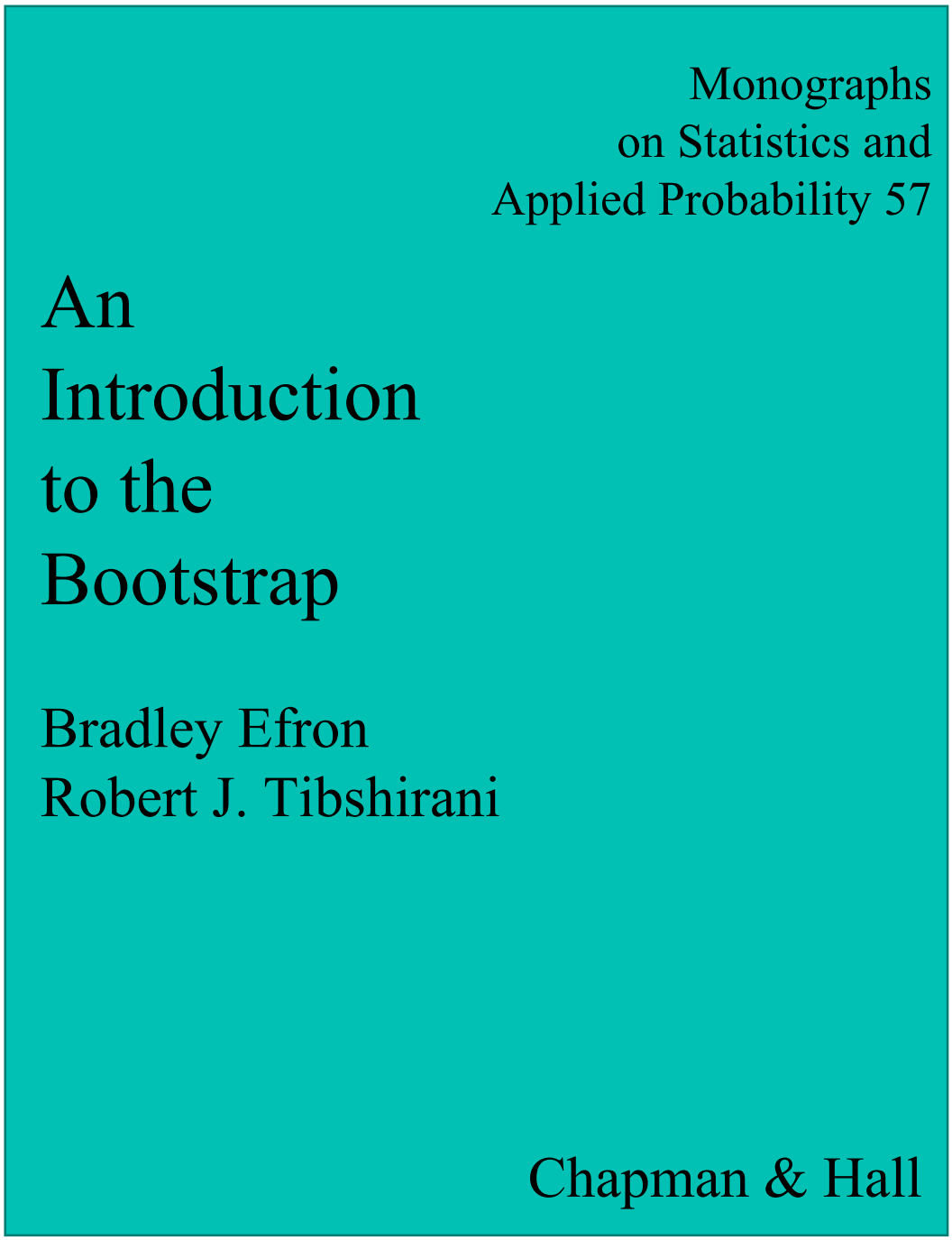
• Erreurs associées aux estimations bootstrap

• Prédiction d'erreurs par l'approche bootstrap

• Bootstrap et images :

➥ détermination des propriétés statistiques (e.g., variance) d'images issues de traitements

Référence recommandée



1993