

Le bootstrap expliqué par l'exemple



Nicolas DESPIEGEL
Lundbeck
<http://www.lundbeck.com>

Olivier DECOURT
Consultant/formateur indépendant
<http://www.od-datamining.com>

2003

Le bootstrap expliqué par l'exemple



1. Les concepts du bootstrap
2. Des variantes adaptées au contexte
3. Comparaison des différentes méthodes
4. Les cas sensibles



Olivier DECOURT
Consultant/formateur indépendant
<http://www.od-datamining.com>

Nicolas DESPIEGEL
Lundbeck
<http://www.lundbeck.com>

2003



1- Les concepts du bootstrap

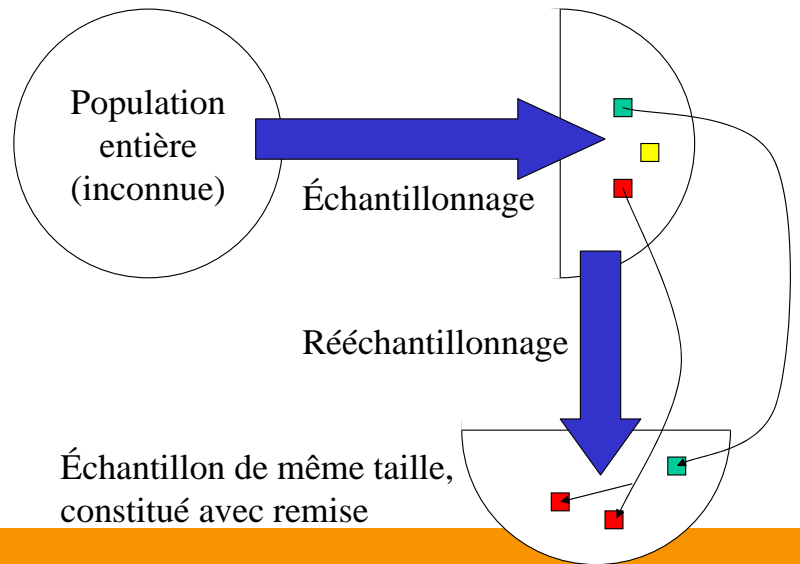
1.1 - Définitions

Rééchantillonnage

Olivier DECOURT
Consultant/formateur indépendant
<http://www.od-datamining.com>

Nicolas DESPIEGEL
Lundbeck
<http://www.lundbeck.com>

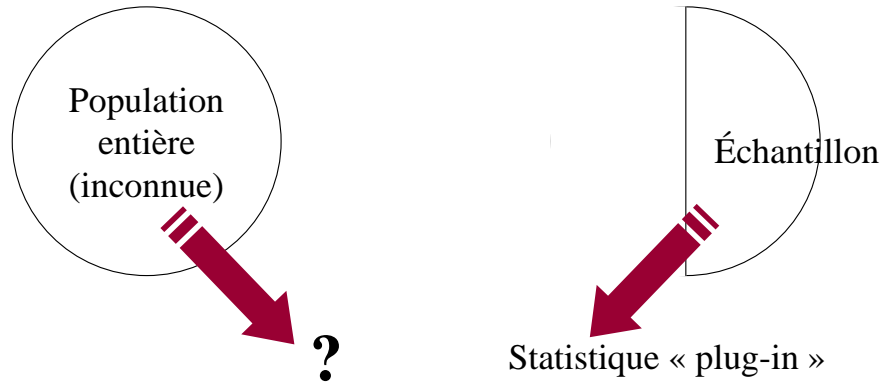
2003



Tout jeu de données peut se considérer comme tiré d'une population inconnue dont on ne voit qu'un extrait. C'est particulièrement vrai dans le cadre des sondages mais cela peut se transposer aisément à une situation générale. L'échantillon peut lui-même jouer le rôle d'une population « mère » : on en tire un nouvel échantillon. C'est le rééchantillonnage. Il s'effectue, dans le cadre du bootstrap, avec remise, de manière à conserver le même nombre de données. On peut en tirer $(2n-1)!/[n!(n-1)!]$ différents. Soit, pour un jeu de 3 individus, 10 échantillons ; pour 4, 35 ; pour 10, 16796...

1.1 - Définitions

Plug-in : calcul d'une statistique et problème d'estimation



Impossible de calculer une statistique sur la population entière ; on lui préfère donc la statistique calculée sur les données de l'échantillon, l'estimateur plug-in. On parle aussi de statistique « empirique ».

1.1 - Définitions

Olivier DECOURT
Consultant/formateur indépendant
<http://www.od-datamining.com>

Nicolas DESPIEGEL
Lundbeck
<http://www.lundbeck.com>

2003

Monte-Carlo (méthode de) :

Mode de calcul d'une quantité inconnue utilisant une suite de nombres *au hasard*

Intérêt : une convergence plus rapide vers la solution qu'une exploration « systématique ».

1.2 – Jackknife et bootstrap

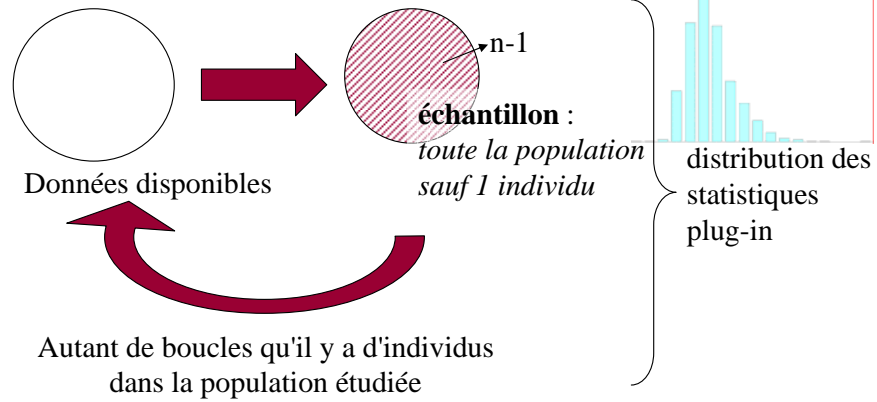
Le jackknife (1/2)

- rééchantillonnage de taille $n-1$ sans remise
- tous les échantillons de taille $n-1$ sont générés
- statistique plug-in calculée sur chaque échantillon de taille $n-1$
- moyenne des plug-in → estimation

Méthode proposée par Quenouille au milieu du XX^{ème} siècle. Analogue aux tests « leave-one-out » ou « validation croisée ». Problèmes : temps de calcul sur un gros échantillon, présence de 2 outliers dans les données => il en reste toujours au moins 1 dans le calcul des plug-in.

1.2 – Jackknife et bootstrap

Le jackknife (2/2)



1.2 – Jackknife et bootstrap

Le bootstrap (1/3)

- rééchantillonnage de taille n avec remise
- génération d'un « grand » nombre d'échantillons
- statistique plug-in calculée sur chaque échantillon
- moyenne des plug-in → estimation

L'idée de base date d'Efron, 1979.

1.2 – Jackknife et bootstrap

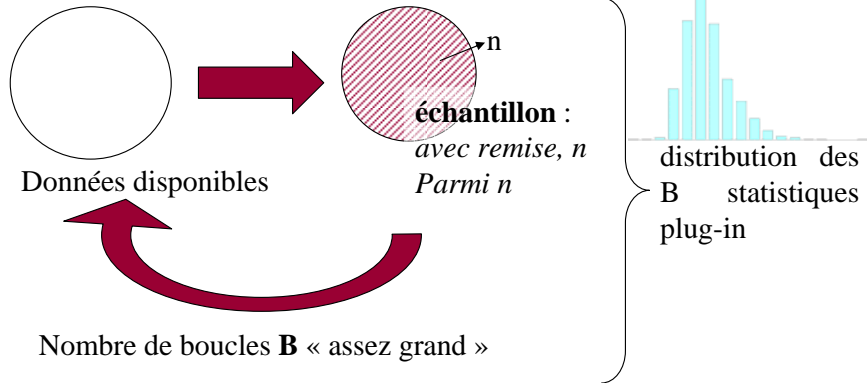
Le bootstrap (2/3)

- idée = reproduire le mécanisme d'obtention des données
- transposition des rôles :
 - les données disponibles jouent le rôle de la population inconnue
 - chaque échantillon bootstrap joue le rôle des données disponibles
- ➔ distribution de la statistique

Reproduire à notre échelle le phénomène qui a permis la création de notre jeu de données = les quantités et les rôles sont transposés en conséquence : données disponibles <-> population « mère » ; stat plug-in sur les données disponibles <-> valeur vraie ; échantillon bootstrap <-> données disponibles ; distribution du plug-in <-> distribution de la statistique.

1.2 – Jackknife et bootstrap

Le bootstrap (3/3)



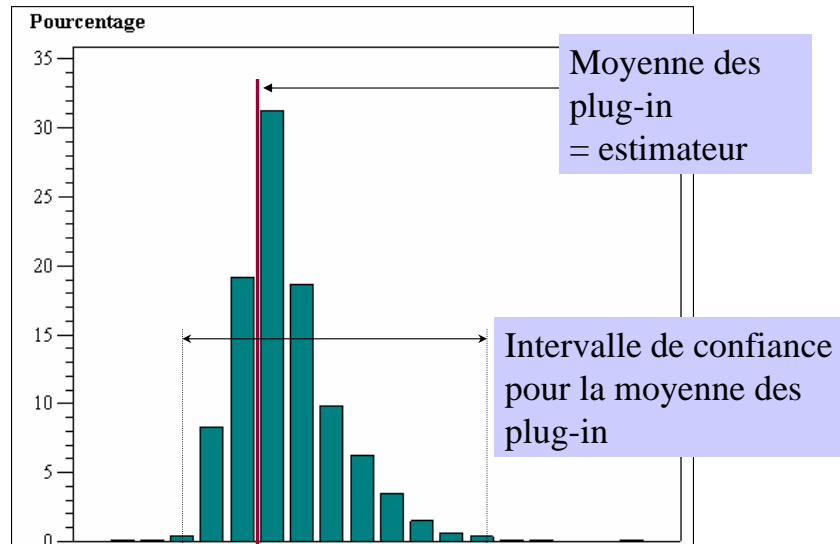
Nombre de boucles : entre 100 et 10000 selon les problèmes et la taille des données disponibles. Les parties 2 et 3 donnent des éléments de choix de ce nombre de boucles.

1.3 – Avantages par rapport à la statistique classique

Olivier DECOURT
 Consultant/formateur indépendant
<http://www.od-datamining.com>

Nicolas DESPIEGEL
 Lundbeck
<http://www.lundbeck.com>

2003



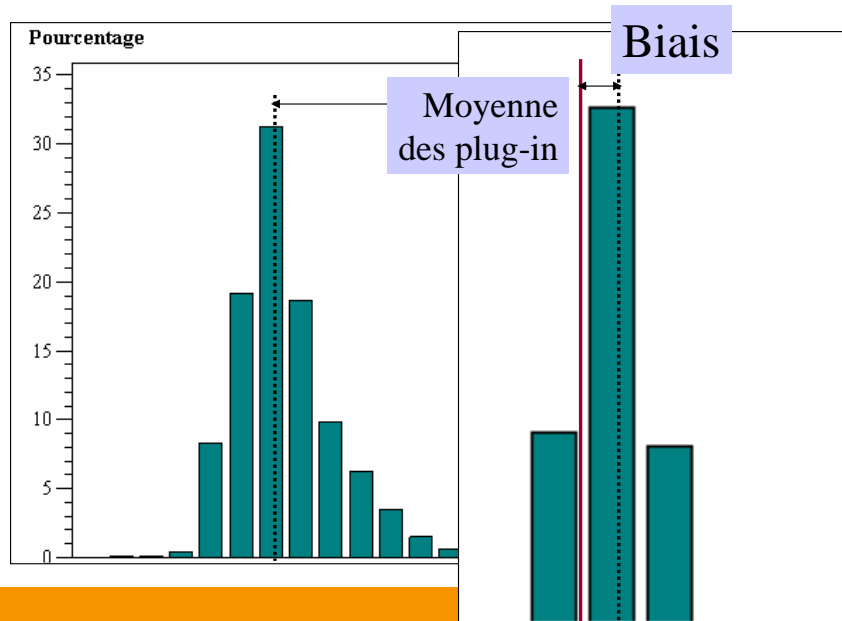
Le bootstrap fournit une DISTRIBUTION et pas seulement un nombre (une moyenne ou une médiane). L'intérêt n'est pas forcément d'avoir une nouvelle estimation : elle n'est pas plus fiable. On peut en revanche calculer un écart-type et/ou un intervalle de confiance pour cette nouvelle estimation (moyenne des plug-in).

1.3 – Avantages par rapport à la statistique classique

Olivier DECOURT
Consultant/formateur indépendant
<http://www.od-datamining.com>

Nicolas DESPIEGEL
Lundbeck
<http://www.lundbeck.com>

2003



Connaître le biais n'implique pas forcément qu'on le corrige. La correction de biais accroît la variance d'une statistique. Il est donc important de connaître le biais **pour avoir un élément de réflexion dans le dilemme biais/variance**. Éventuellement, le bootstrap peut effectivement conduire à une correction de biais.



Olivier DECOURT
Consultant/formateur indépendant
<http://www.od-datamining.com>

Nicolas DESPIEGEL
Lundbeck
<http://www.lundbeck.com>

2003

1.3 – Avantages par rapport à la statistique classique

La fin des problèmes de resubstitution

- Comment calculer un taux de bien classés par un modèle décisionnel ?
- Resubstitution ou validation croisée
→ optimiste
- Bootstrap → quantifie l'optimisme (biais) et le corrige

Pas toujours assez de données pour séparer un jeu de test (cas pharma : courant ; Data Mining : arrive sur des populations très spéciales, CB Gold par exemple).
→ Solution de la resubstitution. Tx d'erreur apparent = tx d'erreur pour les resubstitutions ; biais = optimisme.



Olivier DECOURT
Consultant/formateur indépendant
<http://www.od-datamining.com>

Nicolas DESPIEGEL
Lundbeck
<http://www.lundbeck.com>

2003



2- Des variantes adaptées au contexte

2.1 – Estimation ponctuelle et intervalle de confiance

Bootstrap T

- calcul direct sur les statistiques plug-in...
 - moyenne
 - écart-type
- formule « habituelle » pour l'intervalle de confiance, avec les quantiles d'une **loi de Student** à $n-1$ degrés de liberté

Nécessite entre 500 et 5000 boucles.

Inconvénients : statistiques asymétriques ? Plages de valeurs licites (cf. r^2) ? Ne respecte pas les transformations (par exemple la transformée pour le R^2).

Avantageq : simple et intuitif ; c'est la manière classique de construction d'un IC. Fonctionne très bien qd on peut utiliser une statistique **pivotale** (dont on peut découpler moyenne et variance).

2.1 – Estimation ponctuelle et intervalle de confiance

Bootstrap percentile

- centrage (optionnel) des statistiques plug-in sur la valeur des données disponibles
- **calcul direct des quantiles** sur la distribution des plug-in

Entre 1000 et 5000 boucles sont nécessaires : l'erreur de couverture est en $O(1/n^{1/2})$ ce qui est très lent. (Le bootstrap T et le Bca sont $O(1/n)$.)

Correction de biais : attention à la sortie de la plage de valeurs licite. De +, corriger le biais augmente la variance. On ne corrige le biais que si la variance est inférieure au tiers du biais au carré ($V < 1/3 B^2$).

2.1 – Estimation ponctuelle et intervalle de confiance

Bootstrap Bc_a

- Calcul de deux quantités :
 - **coefficient d'accélération** (proportionnel au skewness)
 - **coefficient de biais** (proportion de plug-in > valeur sur les données disponibles)
- Détermination des quantiles de plug-in qui constituent l'intervalle de confiance

Converge en 200 à 500 boucles seulement (sa probabilité de couverture est $O(1/n)$); problème : détermination des deux coefficients a_0 et z , souvent par Jackknife. Convient plutôt pour déterminer des IC fiables sur de petits volumes de données. Prend en compte les asymétries de la statistique étudiée (ce que ne fait pas le Percentile).

2.2 – Classification et optimisme

Bootstrap de l'optimisme

- calcul du taux d'erreur avec les données disponibles comme jeu de test
- estimateur bootstrap de l'optimisme
- appliqué au taux d'erreur apparent (par resubstitution)

Converge lentement ; nécessite environ 1000 boucles pour bien fonctionner.
Cependant, sa probabilité de couverture est $O(1/n)$.

2.2 – Classification et optimisme

Olivier DECOURT
 Consultant/formateur indépendant
<http://www.od-datamining.com>

Nicolas DESPIEGEL
 Lundbeck
<http://www.lundbeck.com>

2003

Bootstrap .632

- calcul du taux d'erreur apparent (par resubstitution)
- calcul du taux d'erreur bootstrap (avec les données disponibles comme jeu de test)
- formule « magique » combinant les deux quantités



Nom vient du coefficient 0,632 de la formule « magique ». Cela correspond à la probabilité limite d'appartenance d'un individu aux échantillons bootstrap quand B tend vers l'infini. Fonctionne correctement dès 200 boucles.

2.3 – Robustesse

Olivier DECOURT
Consultant/formateur indépendant
<http://www.od-datamining.com>

Nicolas DESPIEGEL
Lundbeck
<http://www.lundbeck.com>

2003

Bagging

- simple bootstrap d'un modèle
- plusieurs variantes selon la finalité du modèle :
 - bootstrap des prédictions
 - bootstrap des coefficients du modèle

Intéressant sur les arbres
de décision
(classificateurs faibles)

Nécessite au plus une centaine de boucles ; le but n'est que de proposer une moyenne de différents modèles délestés de certains outliers

2.3 – Robustesse

Olivier DECOURT
Consultant/formateur indépendant
<http://www.od-datamining.com>

Nicolas DESPIEGEL
Lundbeck
<http://www.lundbeck.com>

2003

Boosting

- variante du bagging
- les rééchantillonnages se font à probabilités inégales
- la probabilité d'inclusion à l'étape b est proportionnelle à l'erreur commise à l'étape $b-1$

A dark blue oval with a subtle gradient and a slight shadow, containing the text 'Convergence ultra-rapide' in white, sans-serif font.

Convergence ultra-rapide

Converge au bout d'une vingtaine de boucles, simple et rapide à utiliser



Olivier DECOURT
Consultant/formateur indépendant
<http://www.od-datamining.com>

Nicolas DESPIEGEL
Lundbeck
<http://www.lundbeck.com>

2003



3- Comparaison des différentes méthodes

3.1 – Intervalle de confiance

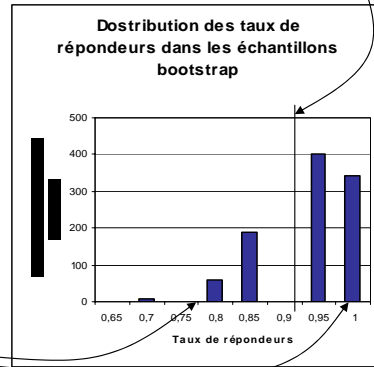
Olivier DECOURT
Consultant/formateur indépendant
<http://www.od-datamining.com>

Nicolas DESPIEGEL
Lundbeck
<http://www.lundbeck.com>

2003

Taux de répondeurs	0,929
Taux de non répondeurs	0,071

Taux sur données réelles



IC à 95%

3.1 – Intervalle de confiance

		IC -	IC +
A (14)	Taux de répondeurs = 0.929		
	Bootstrap-t	0,8	1,058
	Bootstrap percentile	0,786	1
	Bootstrap Bca	0,786	1
B (19)	Taux de répondeurs = 0.737		
	Bootstrap-t	0,541	0,937
	Bootstrap percentile	0,526	0,895
	Bootstrap Bca	0,526	0,947

Bootstrap-t : IC+ dépasse la valeur 1 pour le traitement A

Bootstrap Bca : intervalle légèrement plus large que le percentile en cas d'asymétrie

3.2 – Classification

		Bien classés (jeu de test)	Écart entre entraînement et test
Modèle de base		69,76 %	7,24 %
Bagging	10	69,84 %	6,29 %
	500	70,01 %	6,37 %
Boosting	10	72,74 %	3,14 %
	500	73,90 %	2,35 %

Gain énorme de robustesse pour le boosting, et gain de performance. Intérêt limité, dans les deux cas, de multiplier les boucles. Une centaine est (largement) un maximum. B = 10 donne déjà d'excellents résultats. Données : assurance auto, données réelles maquillées. Modèle : analyse discriminante à deux groupes.



Olivier DECOURT
Consultant/formateur indépendant
<http://www.od-datamining.com>

Nicolas DESPIEGEL
Lundbeck
<http://www.lundbeck.com>

2003



4- Les cas sensibles

4.1 – Où il n’y a rien à gagner

Estimation d’une moyenne

Pour une infinité de tirage, la
moyenne des échantillons
« bootstrap » → moyenne réelle

Intervalle de confiance autour de la moyenne

Normalité quasiment assurée
Résultats connus pour la variance de
la statistique

Moyenne : estimateur sans biais

Cas intéressants :

- Non normalité des données
- Très faible taille de l’échantillon

4.2 – Où il y a tout à perdre

2 affirmations dangereuses

- Pas d'hypothèses sur les données
- Adapté pour les petits échantillons

	Moyenne = 50,6 euros	IC -	IC +
(124)	Bootstrap-t	-10,9	112,1
	Bootstrap percentile	0	120,7
	Bootstrap Bca	8,6	160,6
(19)	Bootstrap-t	-151,4	486,7
	Bootstrap percentile	0	517,9
	Bootstrap Bca	0	863,2

Intervalle de confiance inutile en terme décisionnel

Problématique : Estimation des consommations en soins hospitaliers
 -Distributions très asymétriques, dites à "queue épaisse"
 -Besoin d'information supplémentaire afin de résoudre le problème

Autre type d'exemple: Bootstrap sur un maximum

4.3 – Gagner en robustesse (mais pas n’importe comment)

- gain de **robustesse** → perte de **performance** (et vice-versa)
- données **homogènes** → meilleur espoir de gain sur les 2 tableaux
- **trop de boucles** dans un bagging ou un boosting → faible intérêt
- bagging et boosting → **arbres de décision**, et **rarement** analyse discriminante, régression logistique, réseaux de neurones

L’intérêt du bagging et du boosting s’avère surtout sur un classificateur faible. Les arbres sont donc leur cible privilégiée. L’apport du boosting à un modèle quelconque peut cependant s’avérer intéressant ; le bagging, lui, est souvent décevant sur autre chose qu’un arbre. Le meilleur moyen de gagner en robustesse est l’homogénéité des données (bon nettoyage, modèles stratifiés).



Olivier DECOURT
Consultant/formateur indépendant
<http://www.od-datamining.com>

Nicolas DESPIEGEL
Lundbeck
<http://www.lundbeck.com>

2003

**B fois sur le métier
remettras ton
ouvrage !**



**Vous pouvez télécharger
les macros SAS sur le site
<http://www.ifrance.com/datamining>**

Conclusion essentielle : ce n'est pas parce que le bootstrap sert dans beaucoup de cas qu'il faut TOUT LE TEMPS faire du bootstrap. L'abus de bootstrap est dangereux pour la santé. SAS est un excellent logiciel pour le bootstrap : il permet d'échantillonner (PROC SURVEYSELECT), de boucler un calcul (%DO), de travailler sur un cumul d'information de manière graphique (PROC GCHART, , BOXPLOT, Graph'N'Go, StatGraph en v9) et quantitative (PROC MEANS, SQL, UNIVARIATE).



Olivier DECOURT
Consultant/formateur indépendant
<http://www.od-datamining.com>

Nicolas DESPIEGEL
Lundbeck
<http://www.lundbeck.com>

2003



Des questions ?

