Techniques et Outils de Rééchantillonnage

Dr. Ir. Epiphane SODJINOU
Agroéconomiste, Biostatisticien

CONTENU

1. Introduction: clarification de quelques concepts
2. Différentes techniques de rééchantillonnage
   1. Bootstrap
   2. Jacknife
   3. Validation croisée

0. Objectif global du cours

- Etudier différentes techniques de rééchantillonnage et la manière dont elles peuvent être réalisées avec le logiciel statistique Stata, R et SPSS
- Utiliser ces techniques pour l'estimation de quelques paramètres et modèles

1. Introduction : Clarification de Quelques Concepts

1.1. Notion de rééchantillonnage

- Technique consistant à partir d'un échantillon initial, qui a été réellement prélevé, et à y prélever un ou plusieurs nouveaux échantillons afin d'estimer un paramètre donné
- L'idée qui sous-tend ces méthodes est d'arriver à approcher la distribution (inconnue) de la population à partir d'un échantillon
- Les méthodes de rééchantillonnage construisent des « populations » hypothétiques dérivées des données observées, chacune pouvant être analysée de la même manière pour voir comment les statistiques dépendent de variations aléatoires plausibles dans les données.

1. Introduction : Clarification de Quelques Concepts

1.2. Historique

- Les premières méthodes de rééchantillonnage remontent aux années 1930 avec les tests de permutation et 1949 avec la technique du jackknife prônée par Quenouille (Chernick, 1999)

- Les techniques de rééchantillonnage ont connu, ces dernières années, un développement non moins important et ont fait l'objet d'une littérature abondante.
  - Parmi les principaux auteurs, on peut citer Efron [1979, 1983, 1986], Efron et Tibshirani [1986, 1993], WU [1986], Breiman [1992], Shao [1993, 1996] et Chernick [1999] notamment.

1. Introduction : Clarification de Quelques Concepts

1.2. Historique

- Développement accéléré surtout par l'accès généralisé à des moyens de calcul de plus en plus puissants.
  - Divers logiciels ont intégré ces différentes techniques : "*resampling stats*" (logiciel de MS-DOS pour les tests de randomisation et de rééchantillonnage), SAS, S ou S-PLUS, Matlab, Stata, etc.

1. Introduction : Clarification de Quelques Concepts

1.3. Utilité des techniques de rééchantillonnage

- Ce sont des outils indispensables dans les statistiques modernes
- Le recours à ces méthodes se justifie par le fait que les méthodes classiques ne permettent pas d'obtenir des solutions exactes à tous les problèmes auxquels l'utilisateur se trouve confronté.
- Exemple:
  - L'utilisation de la méthode des moindres carrés (dans le cas de l'ajustement des équations de régression linéaire) suppose que différentes conditions d'application concernant le modèle et les données sont remplies.
  - Le calcul des intervalles de confiance des paramètres tels que la moyenne, la médiane et la variance, suppose que la population-parent est normale et que l'échantillon est aléatoire et simple.
  - Le non-respect de ces conditions peut mettre en question les résultats de l'inférence statistique, suite à une modification des distributions d'échantillonnage des estimateurs qui peut rendre incorrect le calcul des intervalles de confiance ou des tests d'hypothèses.

1. Introduction : Clarification de Quelques Concepts

1.3. Utilité des techniques de rééchantillonnage

- Permet de
  - Tester des hypothèses
  - Calculer des intervalles de confiance pour la plupart des types de données, même celles qui ne peuvent pas être analysées avec des formules.
  - Résolution d'autres problèmes inférentiels
- Des recherches ont montré que les étudiants qui apprennent le rééchantillonnage
  - sont plus enclins à apprendre et appliquer la statistique
  - Estiment que le rééchantillonnage est la façon dont les statistiques devraient être enseignées

1. Introduction : Clarification de Quelques Concepts

1.4. Principe

- Les méthodes de rééchantillonnage sont une extension naturelle de la simulation
- L'analyste utilise un ordinateur pour générer un grand nombre d'échantillons simulés, puis analyse et résume les tendances de ces échantillons.
- La principale différence est que l'analyste commence par les données observées au lieu d'une distribution de probabilité théorique.
- Dans les méthodes de rééchantillonnage, le chercheur ne connaît ni ne contrôle le processus de génération des données (PGD)

1. Introduction : Clarification de Quelques Concepts

1.4. Principe

- Les méthodes de rééchantillonnage commencent par l'hypothèse qu'il existe un PDG de population qui reste non observé, mais que ce PGD a produit le seul échantillon de données observées dont dispose un chercheur
- L'analyste imite ensuite le processus "dans des échantillons répétés" qui pilote les simulations en produisant de nouveaux "échantillons" de données qui consistent en différents mélanges de cas dans l'échantillon d'origine
- Ce processus est répété plusieurs fois pour produire plusieurs nouveaux « échantillons » simulés.

1. Introduction : Clarification de Quelques Concepts

1.4. Principe

- L'hypothèse fondamentale est que toutes les informations sur le PGD contenues dans l'échantillon original de données sont également contenues dans la distribution de ces échantillons simulés
- Si tel est le cas, le rééchantillonnage à partir du seul échantillon observé équivaut à générer des échantillons aléatoires entièrement nouveaux à partir de PGD de la population

1. Introduction : Clarification de Quelques Concepts

1.4. Principe

- Les méthodes de rééchantillonnage peuvent être paramétriques ou non paramétriques
- Dans les deux cas, mais surtout dans le cas non paramétrique, ils sont utiles car ils permettent à l'analyste d'assouplir une ou plusieurs hypothèses associées à un estimateur statistique.
- Les erreurs standard des modèles de régression, par exemple, reposent généralement sur le théorème central limite ou la normalité asymptotique des estimations de ML
- Les méthodes de rééchantillonnage peuvent être facilement adaptées à la non-indépendance entre les observations.

2. Différentes techniques de rééchantillonnage

2.0. Différents types de rééchantillonnage

- Différents types de rééchantillonnage (Efron et Tibshirani, 1993 ; Chernick, 1999)
  - Bootstrap
  - Jackknife
  - Validation croisée
  - Tests de randomisation ou de rerandomisation,
  - Tests de permutation ou d'autocomparaison

2.1. Bootstrap

Méthode du Bootstrap

- Méthode proposée par EFRON [1979]
- Particulièrement adaptée à des situations où la distribution de la population-parent n'est pas connue,
  - mais peuvent aussi être utilisées dans le cas où celle-ci serait connue
- Principe:
  - Considérons un échantillon initial x, de n observations (x_1, x_2, …, x_n), prélevé de manière aléatoire et simple dans une population.
  - Considérons le cas le plus simple où ces observations concernent une seule variable.

2.1. Bootstrap

Méthode du Bootstrap

- Principe:
  - Effectuer, au sein de l'échantillon initial *x*, une série d'échantillonnages aléatoires, simples, indépendants les uns des autres, de même effectif *n* et avec remise
  - On calcule pour chacun de ces échantillons, la valeur du paramètre considéré.
  - Ces échantillons successifs seront notés :

$$X_1^*\ ,\ X_2^*\ ,\ \ldots,\ X_k^*\ ,\ \ldots,\ X_B^*\ ,$$

B étant le nombre d'échantillons prélevés.

2.1. Bootstrap

Méthode du Bootstrap

- Les échantillons ainsi obtenus, sont appelés des échantillons du bootstrap (*bootstrap samples*)
- Pour un échantillon donné x_k^* du bootstrap, une observation quelconque x_i, de l'échantillon initial, peut apparaître 1, une, deux, …, ou *n* fois
- On peut, ainsi, définir pour l'ensemble des *B* échantillons, les proportions d'apparition P_i^* de chacune des observations initiales x_i.
- On a: $P_i^* = \frac{n_i}{nB}$ …
  - avec n_i le nombre de fois que l'observation i a été prélevée pour l'ensemble des *B* échantillons
- Ces proportions interviennent dans certaines estimations
- Des méthodes de rééchantillonnage assurant l'égalité de ces proportions existent et conduisent au « rééchantillonnage balancé »

2.2. Jackkknife

Principe

- Considérons un échantillon initial x , de *n* observations $(x_1, x_2, \ldots, x_n)$
- Dans le cas le plus simple, la méthode du jackknife consiste à éliminer à tour de rôle chacune des *n* observations, en formant de cette manière *n* sous-échantillons de *n-1* observations
- On calcule ensuite, le paramètre considéré pour chacun de ces *n* sous-échantillons.
- Cette méthode peut être étendue, en éliminant à tour de rôle tous les couples de deux individus, de trois, ou toute autre partie de l'échantillon [DAGNELIE, 1998].

2.2. Jackkknife

Avantages

- Le jackknife a des avantages, comme sa nature non paramétrique qui le rend robuste à certaines violations d'hypothèses
- Comme d'autres méthodes de rééchantillonnage, elle peut également être adaptée à de nombreuses structures de données différentes, telles que les données en cluster, dans lesquelles des groupes entiers d'observations (plutôt qu'une seule observation) sont supprimés à chaque itération
- Il est également bon pour détecter les valeurs aberrantes et/ou les cas influents dans les données.
  - En effet, sa procédure de non-participation est similaire au *D de Cook*, qui est souvent utilisé pour détecter les valeurs aberrantes dans les modèles de régression linéaire (Cook, 1977).

2.2. Jackkknife

Limites

- Jackknife ne fonctionne pas aussi bien si la statistique d'intérêt ne change pas « en douceur » d'une répétition à l'autre
  - Par exemple, le jackknife sous-estimera l'erreur type de la médiane dans de nombreux cas parce que la médiane n'est pas une statistique lisse (voir Rizzo, 2008)
  - Dans de tels cas, il est nécessaire d'omettre plus d'une observation à la fois (Efron & Tibshirani, 1993)
- Jackknife peut être problématique dans les petits échantillons car la taille de l'échantillon dicte le nombre de répétitions/rééchantillonnages possibles.
  - Cette limitation n'est pas rencontrée par la méthode de rééchantillonnage la plus polyvalente et peut-être la plus courante : le bootstrap.

2.3. Tests de permutation et de randomisation

- Tests de permutation
  - Forme la plus ancienne de méthodes de rééchantillonnage, remontant aux travaux de Ronald A. Fisher dans les années 1930 (Fisher, 1935 ; Pitman, 1937)
  - Sont généralement utilisés pour tester l'hypothèse nulle selon laquelle l'effet d'un traitement est nul
  - Plutôt que de supposer une forme particulière pour la distribution nulle, l'analyste utilise les données observées pour en créer une
  - Cela se fait en mélangeant au hasard l'échantillon plusieurs fois, en créant de nouveaux échantillons qui "cassent" la relation dans l'échantillon observé à chaque fois

2.3. Tests de permutation et de randomisation

- Ensuite,
  - la statistique d'intérêt est calculée dans chaque échantillon remanié
- Enfin,
  - l'estimation de l'échantillon original est comparée à la distribution des estimations des échantillons remaniés pour évaluer la différence entre l'estimation observée et le remaniement aléatoire
- Si chaque combinaison de remaniement est calculée, la procédure est appelée **test de permutation** ou **test exact**
- Une autre option consiste à effectuer un « grand » nombre de remaniements, auquel cas cela s'appelle un **test de randomisation**

2.3. Tests de permutation et de randomisation

Principe

- Supposons que vous disposiez d'un échantillon d'individus
- Un sous-ensemble d'entre eux a reçu un traitement (t_1) alors que les autres n'en ont pas reçu (t_0)
- La base de données indique si oui ou non chaque individu de l'échantillon a reçu le traitement, et donne aussi pour chaque individu une variable dépendante d'intérêt, Y
- Supposons que vous souhaitiez enregistrer si les moyennes de Y diffèrent pour les deux groupes (t_1 et t_0)
- Un test de permutation remanierait les données en gardant la valeur observée de Y de chaque individu inchangée, mais en attribuant au hasard t_1 ou t_0 aux individus

2.3. Tests de permutation et de randomisation

Principe

- La différence des moyennes de Y entre ces deux « groupes » serait calculée et enregistrée
- Ce processus de remaniement qui a été enregistré comme recevant le traitement ou non serait répété plusieurs fois, la différence dans les moyennes de Y étant enregistrée à chaque fois
- La différence réelle observée dans les moyennes de Y pour l'échantillon original serait alors comparée à la distribution de ces différences simulées dans les moyennes qui ont émergé du hasard

2.3. Tests de permutation et de randomisation

Principe

- L'objectif est d'évaluer si la différence de moyenne observée dans l'échantillon réel diffère suffisamment de la distribution de celles générées aléatoirement pour que le chercheur conclue que le traitement a un impact sur Y
- Un test de randomisation est un test de permutation dans lequel un grand nombre de permutations possibles sont assemblées et analysées au lieu de chacune d'entre elles.

2.3. Tests de permutation et de randomisation

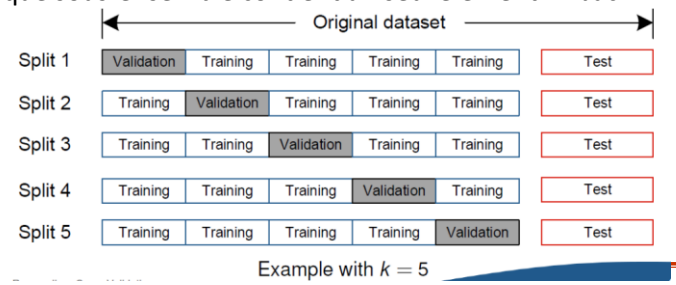
Hypothèse de base

- Les tests de permutation et de randomisation supposent l'interchangeabilité, ce qui signifie que les résultats observés entre les individus proviennent de la même distribution, quelle que soit la ou les valeurs de la ou des variables indépendantes (Kennedy, 1995)

- Il s'agit d'une hypothèse plus faible que l'hypothèse iid (independent and identically distributed), qui inclut également la notion d'indépendance

2.4. Validation croisée (Cross-Validation)

- Diviser au hasard l'ensemble de données en k sous-échantillons de taille égale
- Ajuster les données en utilisant k-1 sous-ensemble
- Faire des prédictions avec le $k^{ième}$ sous-ensemble qui avait été laissé de côté
- Mesurer la performance
- Répétez cette opération k fois de sorte que chaque sous-ensemble devienne un ensemble de validation
- Si chaque sous-ensemble contient un seul élément ➔ Jackknife



3. A savoir

- Les méthodes de rééchantillonnage sont similaires à la simulation en ce sens qu'elles utilisent un processus itératif pour résumer les données.
- La principale différence est qu'ils s'appuient sur l'échantillon de données observé plutôt que sur une distribution théorique
- Cette base empirique confère généralement aux méthodes de rééchantillonnage une robustesse aux violations d'hypothèses

3. A savoir

- Chaque technique de rééchantillonnage est adaptée à des situations spécifiques
- Tests de permutation et de randomisation
  - Généralement meilleurs pour les données expérimentales où il existe une variable de traitement claire et une hypothèse nulle d'absence d'effet.
  - Ont été étendus au cadre de régression multiple (Erikson et al., 2010), mais nous ne le recommandons pas car il ne permet pas d'estimer la matrice de variance-covariance complète des estimations des coefficients
- Jackknife
  - Bonne option lorsque l'analyste s'inquiète de l'influence indue par une donnée particulière
  - Cependant, il fonctionne mal dans de petits échantillons et lorsque la statistique d'intérêt n'est pas lisse.

3. A savoir

- Bootstrap
  - Recommandé dans la plupart des cas, car il est flexible et robuste pour de nombreux types de données différents.
  - Permet d'estimer la matrice de covariance complète dans un modèle de régression multiple tout en fonctionnant généralement bien dans de petits échantillons
  - Son lien étroit avec la conceptualisation de la simulation d'échantillons répétés rend la méthode intuitivement attrayante
  - Même s'il ne s'agit pas d'une panacée, le bootstrap est un outil très utile pour les spécialistes des sciences sociales appliquées.