Traitement des valeurs atypiques d'une enquête par winsorization - application aux enquêtes sectorielles annuelles

Thomas Deroyon - DMS Insee
thomas.deroyon@insee.fr

1er avril 2015

Plan

Unité atypique - une définition

Définition :

- une unité est atypique si elle a des réponses très différentes des réponses des unités de l'échantillon ayant le même poids
- Exemple - dans un sondage stratifié à un degré avec SAS dans les strates, une unité est atypique si ses réponses diffèrent fortement des réponses des autres unités de la strate

Exemple simple : sondage stratifié à un degré

- Population U, divisée en H strates U_h de taille N_h
- Sélection d'un échantillon s_h dans la strate U_h par sondage aléatoire simple de n_h unités dans chaque strate
- **Estimateur d'Horvitz-Thompson** du total de la variable d'intérêt X : $\hat{X} = \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{i \in s_h} X_i$
- **Variance** : $\mathbb{V}(\hat{X}) = \sum_{h=1}^{H} N_h^2 (1 - \frac{n_h}{N_h}) \frac{S_h^2(X)}{n_h}$
  - avec $S_h^2(X) = \frac{1}{N_h - 1} \sum_{i \in U_h} (X_i - \bar{X}_h)^2$
  - et $\bar{X}_h$ moyenne de X dans la strate U_h

Unité atypique $\Rightarrow X_i$ très différent de $\bar{X}_h \Rightarrow S_h^2(X)$ élevé $\Rightarrow$ Variance de $\hat{X}$ élevée

Deux types d'unités atypiques

- UNITÉS ATYPIQUES NON REPRÉSENTATIVES : l'unité ne peut représenter qu'elle même
  - erreur de déclaration
  - caractéristique très atypique : participation à une restructuration ...
  - identification et traitement à la phase d'apurement (*data editing*)
- UNITÉS ATYPIQUES REPRÉSENTATIVES : l'unité diffère des autres unités de sa strate, mais on ne peut pas supposer qu'elle ne représente qu'elle même. Identification et traitement :
  - modèles de mélange
  - méthodes de biais conditionnel
  - winsorization

Principe

Cadre : sondage stratifié à un degré avec sondage aléatoire simple dans chaque strate

Principe :

- choix d'une variable d'intérêt X
- définition de seuils K_h dans chaque strate
- création d'une variable winsorizée X^w obtenue en rabotant les valeurs de X qui dépassent les seuils dans chaque strate
- deux types de winsorization :

$$X^w = \begin{cases} X \text{ si } X < K_h \\ \textbf{Type I :} \ K_h \text{ si } X > K_h \\ \textbf{Type II :} \ \frac{n_h}{N_h} X + (1 - \frac{n_h}{N_h}) K_h \text{ si } X > K_h \end{cases}$$

Arbitrage biais variance

Estimateur winsorizé : l'estimateur winsorizé du total de X est l'estimateur d'Horvitz-Thompson du total de la variable winsorizée

$$\hat{X}^w = \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{i \in s_h} X_i^w$$

- estimateur biaisé du total de X
- variance plus faible car la variance empirique $S_h^2(X^w)$ de X^w est plus faible que $S_h^2(X)$ dans chaque strate
- $\Rightarrow$ arbitrage **biais - variance**
- **paramètre central** : choix des seuils $K_h \Rightarrow$ méthode de Kokic et Bell

Principe

HYPOTHÈSES :

- sondage stratifié à un degré avec sondage aléatoire simple dans chaque strate
- variable winsorizée X à valeurs positives ou nulles
- dans une strate, toutes les valeurs de X sont issues de la même loi, d'espérance μ_h
- les seuils K_h sont indépendants de l'échantillon auquel ils sont appliqués

OBJECTIF :

calcul de seuils qui nous protègent contre les unités atypiques

- quelles que soient les valeurs de X dans la population
- quel que soit l'échantillon sélectionné

Formules à l'optimum 1/2

$\Rightarrow$ Calcul des seuils qui minimisent l'erreur quadratique moyenne de l'estimateur winsorizé sous

- l'aléa résultant du plan de sondage
- la distribution de X dans chaque strate

RÉSULTATS :

- à l'optimum
- et asymptotiquement (quand $N_h \to +\infty$ et $n_h \to +\infty$)
- les seuils K_h^* et le biais de l'estimateur winsorizé B^* vérifient :

$$K_h^* = \mu_h - \frac{B^*}{\frac{N_h}{n_h} - 1}$$

- $K_h \to +\infty$ quand $\frac{n_h}{N_h} \to 1$
- K_h est proche de μ_h quand le taux de sondage est très faible

Formules à l'optimum 2/2

RÉSULTATS :

- le biais de l'estimateur winsorizé B^* est le point où la fonction F s'annule avec :

$$F(B) = -B\,[1 + \sum_{h=1}^{H} n_h\, E_h(J_h^*)] - \sum_{h=1}^{H} n_h\, E_h(X_h^* J_h^*) \text{ avec}$$

  - E_h espérance sous la loi de X dans la strate h
  - $X_h^* = (\frac{N_h}{n_h} - 1)(X_h - \mu_h)$
  - J_h^* indicatrice que l'unité est winsorizée $(X > K_h)$, aussi égale à l'indicatrice que X_h^* dépasse $-B$
  - $E_h(J_h^*)$ probabilité qu'une unité de la strate soit winsorizée
  - $E_h(J_h^* X_h^*)$ moyenne de la variable égale à X_h^* sur les unités winsorizées et 0 ailleurs

En pratique

(Nouvelle) Hypothèse : nous disposons d'observations $\tilde{X}$ de X dans chaque strate **indépendantes de l'échantillon winsorizé**

- estimation de μ_h dans chaque strate par la moyenne empirique des $\tilde{X}$ dans la strate
- calcul des $\tilde{X}_h^*$
- pour chaque valeur possible du biais B, estimation de $E_h(J_h^*)$ par la part des valeurs de $\tilde{X}$ supérieure à $-B$ dans la strate
- pour chaque valeur possible du biais B, estimation de $E_h(X_h^* J_h^*)$ par la moyenne de la variable égale à $\tilde{X}_h^*$ si $\tilde{X}_h^* > -B$ et 0 sinon
- estimation du zéro de la fonction $F \Rightarrow$ estimation du biais optimal B^*
- estimation des seuils optimaux K_h^* par $K_h^* = \hat{\mu}_h - \frac{-\hat{B}^*}{\frac{N_h}{n_h}-1}$

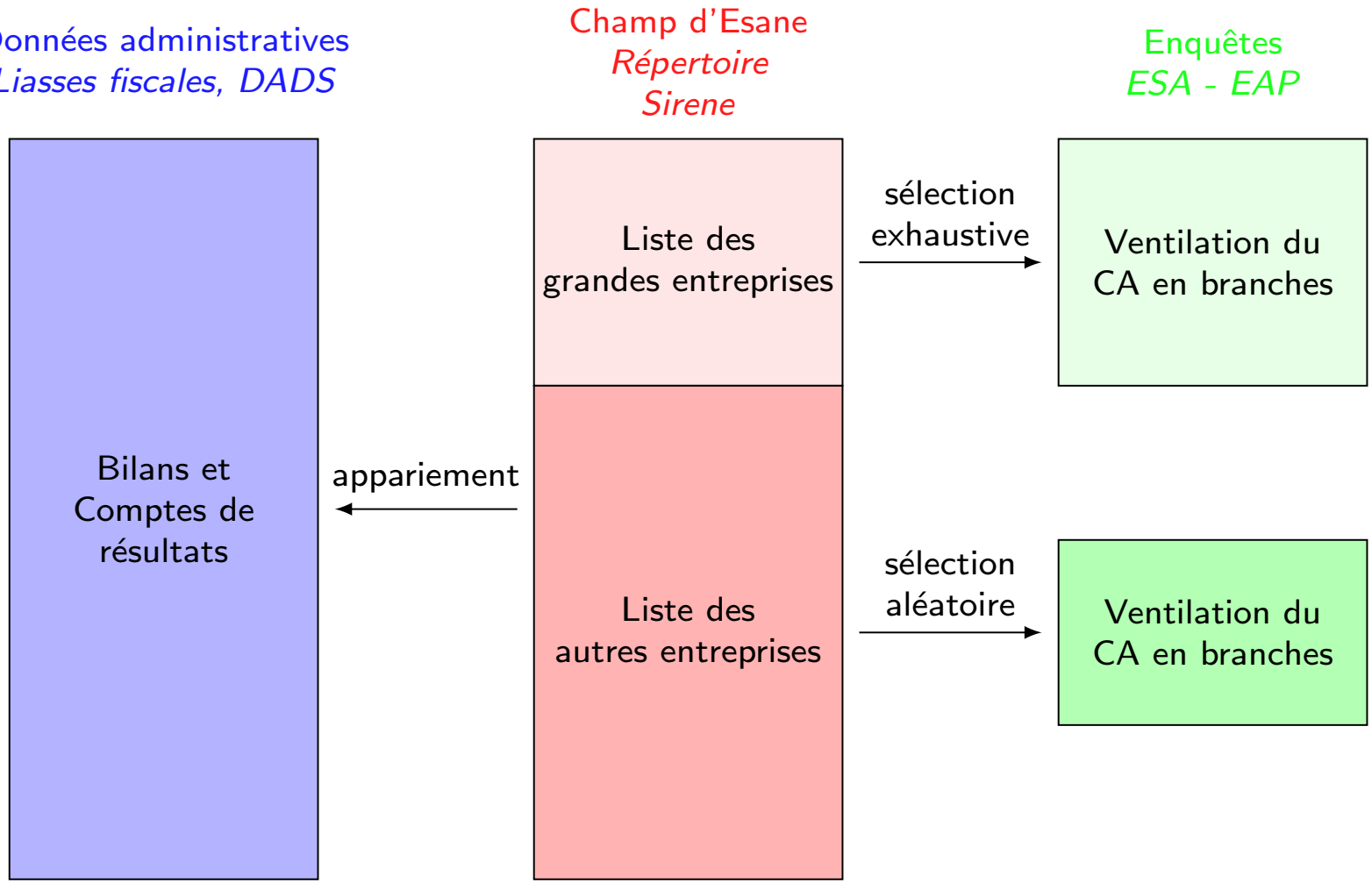
$\Rightarrow$ **Problème** : trouver des valeurs de X sur des unités indépendantes de l'échantillon dans chaque strate

Objectifs

ESANE : Elaboration des Statistiques ANnuelles d'Entreprise
$\Rightarrow$ estimation des statistiques structurelles d'entreprise
Objectifs :

- **comptes par secteur** : estimation du total des variables de comptes de résultat ou de bilan des entreprises par secteur (ensemble des entreprises ayant la même activité principale)
- **ventilation du chiffre d'affaires par branche** : chaque entreprise a d'autres activités que son activité principale $\Rightarrow$ quantifier le chiffre d'affaires par activité, pour estimer les comptes de branche (fonctions de production) dans la comptabilité nationale

Les données


Données administratives
Liasses fiscales, DADS
Champ d'Esane
Répertoire Sirene
Enquêtes
ESA - EAP
Bilans et Comptes de résultats
appariement
Liste des grandes entreprises
Liste des autres entreprises
sélection exhaustive
sélection aléatoire
Ventilation du CA en branches
Ventilation du CA en branches

Plan de sondage

Plan de sondage stratifié à un degré sur

- le secteur (sous-classe de la NAF rev.2)
- effectif
- région

Strates :

- exhaustives : 80 000 entreprises chaque année environ
- non exhaustives : 70 000 entreprises échantillonnées chaque année, tirées parmi 2 millions d'entreprises (taux de sondage par strate variant de 1/3 à 1/400)

Hypothèses simplificatrices sur le plan de sondage

Problème :

- Renouvellement de l'échantillon - chaque année, la moitié de l'échantillon tiré dans les strates non exhaustives l'année précédente est conservée dans l'échantillon
- Winsorization sur les répondants - tenir compte de la non-réponse

Hypothèses :

1. pas de prise en compte du renouvellement
2. sélection des répondants assimilée à un sondage stratifié dans les strates de tirage initiales

Quelle(s) variable(s) winsorizer ?

Estimateurs composites d'Esane combinent données administratives et données d'enquête
Ils font intervenir (entre autres) :

- des sommes pondérées des variables déclarées dans l'enquête (CA par branche)
- des sommes pondérées sur les répondantes aux enquêtes des variables déclarées dans la liasse fiscale

$\Rightarrow$ possibilité d'unités atypiques sur les variables de l'enquête et de la liasse fiscale
Problème : les variables sont liées entre elles par de nombreuses relations comptables
Solution : winsorization sur le chiffre d'affaires fiscal CA - toutes les autres variables sont corrigées par $\frac{CA^w}{CA}$

Quelles données sont disponibles ? 1/2

Nécessité de disposer de

- chiffres d'affaires fiscaux par strate
- sur un ensemble d'entreprises indépendant de l'échantillon de l'enquête

Candidats possibles :

- EAE 2007 : enquêtes existant avant Esane
  - demandaient aux entreprises leurs comptes de résultat et bilans
  - échantillon indépendant de celui des ESA et des EAE
  - utilisées pour calculer les seuils utilisés de 2008 à 2012
  - Problème : ne couvraient pas l'industrie + nécessité de les actualiser
- Edition précédente de l'enquête : pour rester dans les hypothèses de la méthode, enquête $N-2$ ou précédente

Quelles données sont disponibles ? 2/2

Candidats possibles :

- Chiffres d'affaires disponibles dans les liasses fiscales de toutes les entreprises de la base de sondage
- Test : chiffres d'affaires fiscaux des entreprises de l'échantillon de l'enquête winsorizée
  - dispositif Esane particulier : on a rarement la variable à winsorizer sur la base de sondage
  - l'utilisation des données de l'enquête à winsorizer est une des seules stratégies possibles en cas d'enquête non répétée
  - $\Rightarrow$ qu'est ce qui se passe quand on ne respecte pas l'hypothèse d'indépendance des seuils de winsorization à l'échantillon

Résultats 1/2

Sur les ESA et EAP 2012 :

- **Seuils base de sondage** : winsorization de 220 entreprises, pour un effet de $-1,1$ milliards d'€ en données non pondérées, -41 milliards d'€ en données pondérées
- **Seuils enquête 2010** : winsorization de 469 entreprises, pour un effet de $-1,3$ milliards d'€ en données non pondérées, -48 milliars d'€ en données pondérées
- **Seuils enquête 2012** : winsorization de 374 entreprises, pour un effet -750 millions d'€ en données non pondérées, -26 milliars d'€ en données pondérées
- **Seuils EAE 2007** : 268 entreprises, pour un effet, pour un effet de $-1,3$ milliards d'€en données non pondérées, -49 milliards d'€en données pondérées
- Seuils enquête 2012 et 2010 : winsorize beaucoup d'unités
- Seuils BDS 2012 : winsorization concentrée sur un nombre restreint d'unités

Résultats 2/2

- Quel que soit le seuil, la winsorization **améliore la précision** des estimateurs
- Pas de grande différence entre les précisions des estimateurs suivant les seuils de winsorization choisis
- **Interprétation** : effet de la winsorization sur la précision se concentre sur un nombre limité d'unités, qui sont fortement winsorizées quels que soient les seuils (les 10 entreprises les plus winsorizées concentrent la moitié de l'effet sur le chiffre d'affaires)

$\Rightarrow$ utilisation des seuils calculés sur la base de sondage
$\Rightarrow$ étude sur simulations de l'utilisation des seuils calculés sur l'enquête winsorizée