Manuela Maia

Estimateur optimal dans l'échantillonnage Indirect utilisant deux bases de sondages

Septième colloque francophone sur les sondages

5 au 7 Novembre

ENSAI | BRUZ | RENNES| FRANCE

GOVERNO DA REPÚBLICA PORTUGUESA

FCT Fundação para a Ciência e a Tecnologia
MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E ENSINO SUPERIOR Portugal

Introduction

Base de sondage imparfaite

- Sous-couverture
- Duplication
- Sur-couverture

Erreur de couverture sur les statistiques

Problèmes d' estimation (calcul des poids de sondage)

Réduire l'impact des erreurs de couverture

Base de Sondage avec sous-couverture

Combiner plusieurs bases

Bases multiples

Sondage Indirect (Lavallée, 1995) – approche alternative à la théorie classique de l'échantillonnage pour traiter le problème des bases de données chevauchantes

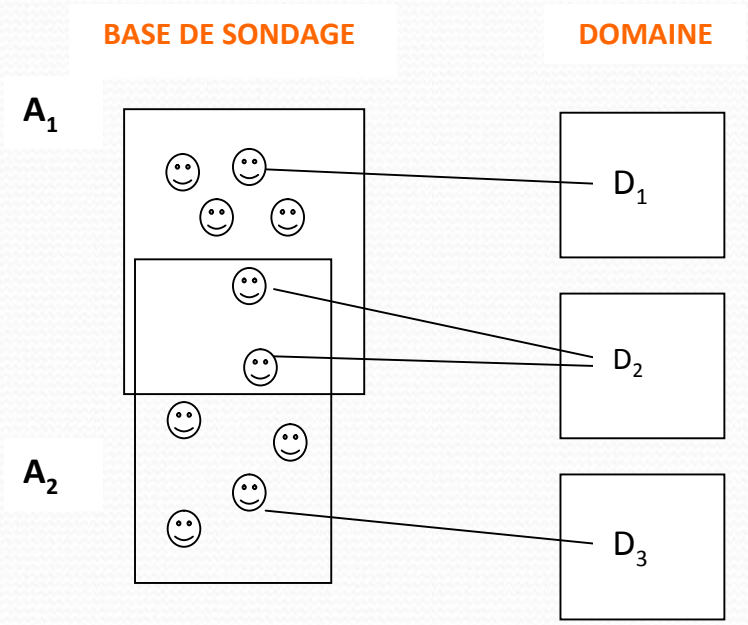
- Nouvelle classe d'estimateurs réunissant des estimateurs pour les bases de sondage multiples et des estimateurs pour le sondage indirect

Un seul estimateur pour tenir compte des bases multiples

- Comparer l'estimateur optimal de Deville et Lavallée avec deux classes d'estimateurs : estimateur "Domain Membership", et estimateur "Unit Multiplicity"

Estimateurs pour bases multiples

Le total Y à estimer sur l'union des Q bases de sondage chevauchantes est exprimé comme une somme sur l'ensemble des 2^Q-1 domaines disjoints.



$D_1 = A_1 \cap \bar{A}_2$
$D_2 = A_1 \cap A_2$
$D_3 = \bar{A}_1 \cap A_2$

Estimateur "Dual frame" du total sur la population proposé par Hartley (1974), basé sur une moyenne pondérée des estimateurs de total sur les domaines :

$$\hat{Y} = \hat{Y}_{D_1}^{A_1} + \lambda\, \hat{Y}_{D_2}^{A_1} + (1-\lambda)\, \hat{Y}_{D_2}^{A_2} + \hat{Y}_{D_3}^{A_2}$$

$0 \leq \lambda \leq 1$ Paramètre à minimiser $V[\hat{Y}(\lambda)]$

$\hat{Y}_{D_1}^{A_1}$ Estimateur du total de D_1

$\hat{Y}_{D_3}^{A_2}$ Estimateur du total de D_2

$\hat{Y}_{D_2}^{A_2}$ Estimateur du total de D_2 avec l'échantillon de A_2

$\hat{Y}_{D_2}^{A_1}$ Estimateur du total de D_1 avec l'échantillon de A_1

L'échantillon sélectionné dans chaque base est utilisé pour produire un estimateur du total dans chaque domaine. On les combine ensuite pour obtenir un seul estimateur pour le total sur la population.

$$\hat{Y} = \sum_{K} \sum_{q \in \mathrm{K}} \sum_{i \in \bigcup_{\mathrm{q}} \mathrm{A_q}} w_i^{(q)} \delta_i(K)\, \mathrm{y_i}$$

Indicateur d'appartenance au domaine $\delta_i(K) = \begin{cases} 1 & ,i \in D_k \\ 0 & ,otherwise \end{cases}$

Les poids $w_i^{(q)}$ doivent être calculés

Approche "Domain Membership"

Approche "Unit multiplicity"

Approche "Domain Membership"

La partition des domaines est définie sur les bases de sondage – il est toujours possible d'identifier correctement à quel domaine appartient chaque unité de l'échantillon

Trois types d'estimateurs, dépendant des poids fixés :

(a) *Estimateur Optimal* $w_{i,opt}^{(q)}$

Présente de bonnes propriétés théoriques - variance minimale (Hartley 1962, 1974; Lund 1968; Fuller et Burmeister 1972) – mais très complexe de façon opérationnelle

(b) Estimateur "*Single Based*" $w_{i,SF}^{(q)}$

Utilise des poids fixés assurant des estimateurs sans biais (Bankier 1986; Kalton et Anderson 1986; Skinner 1991; Skinner, Holmes et Holt 1994), mais qui sont moins efficaces que pour l'estimateur optimal (Lohr and Rao 2000)

(c) Estimateur du *Pseudo Maximum de Vraisemblance* $w_{i,PML}^{(q)}$

Etend l'applicabilité de l'estimateur optimal, en améliorant son eficaicté par rapport à l'estimateur "single based" (Skinner et Rao 1996; Lohr et Rao 2000)

Approche “Multiplicity Unit”

Estimateurs basés sur le concept d’unité multiple, qui reflète le nombre de base de sondage auxquelles chaque élément de l’échantillon appartient (Mecatti 2007) Casady(1980) and Sirken (2004)

$$\hat{Y}_M = \sum_{q=1}^{Q} \sum_{i \in \mathbf{s}_q} w_i^{(q)} \, \mathrm{y_i} \, m_i^{-1}$$

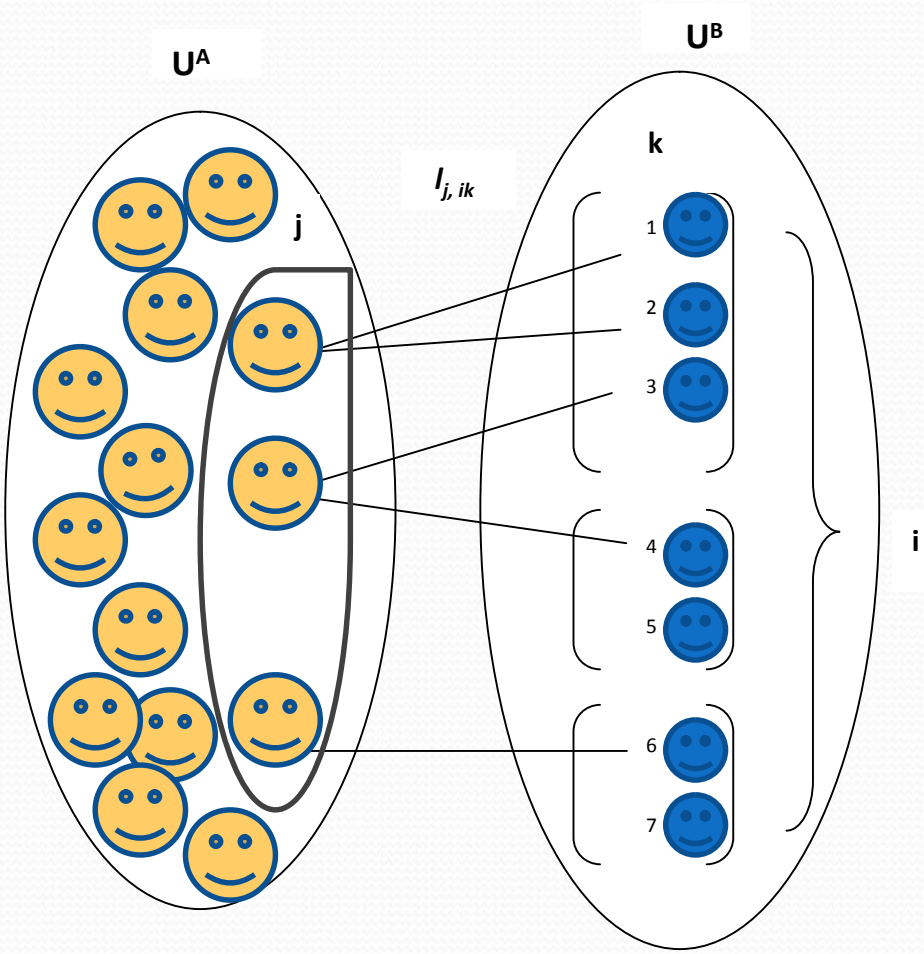
Q est le nombre de bases,

$$\mathrm{m_i} = \sum_{q} \delta_i^{(A_q)} \qquad \delta_i^{(A_q)} = \begin{cases} 1 & \text{if} \quad \mathrm{i} \in \mathrm{A_q} \\ 0 & \text{if} \quad \mathrm{i} \notin \mathrm{A_q} \end{cases}$$

est le nombre de bases dans lesquelles chaque unité est incluse parmi les bases utilisées pour l’enquête

Mecatti (2007) donne des arguments pour appliquer ces estimateurs dans les enquêtes avec plus de deux bases de sondage.

Sondage Indirect



U^A avec M^A unités où nous tirons un échantillon s_A avec m_A éléments

U^B contient M^B éléments, divisés en N grappes, chacune avec M_i^B éléments

$I_{j,ik}$ - liens entre les unités $j \in U^A$ et les éléments *k* de la $i^{ème}$ grappe de U^B pour calculer le poids de chaque élément de l'échantillon

Sondage Indirect

Utilise les liens entre la population cible U^B et la base de sondage U^A pour produire un estimateur d'un paramètre de U^B, quand la base de sondage ne correspond pas parfaitement à U^B

Méthode de Partage des Poids Généralisés (GWSM) (Lavallée, 1995)

Fournit des poids d' estimation pour les unités enquêtées de U^B

Moyenne des poids d'échantillonnage de U^A

"Méthode de partage des poids" Ernest (1989)

"Echantillonnage de réseaux" Thompson (1992)

"Sondage adaptatif de grappes" Thompson and Seber (1996)

Résultats

Pour le total sur la population dans le cas d'un Sondage Indirect nous pouvons écrire le :

Estimateur "Domain Membership"

$$\hat{Y}_{DM} = \sum_{j \in A_1} \frac{t_j^{A_1} z_j(\lambda)}{\pi_j^{A_1}} y_j + \sum_{j \in A_2} \frac{t_j^{A_2} x_j(\lambda)}{\pi_j^{A_2}} y_j$$

$\pi_j^{A_q}$ représente la probabilité de sélection de l'unité j dans la q^{ème} base

$$t_j^{A_q}(\theta) = \begin{cases} 1 & if \quad j \in A_q \\ 0 & if \quad j \notin A_q \end{cases}, \quad q = 1,2$$

indicateurs des bases de sondage

$$z_j(\lambda) = \begin{cases} 1 & if \quad j \in D_1 \\ \lambda & if \quad j \in D_2 \end{cases}$$

$$x_j(\lambda) = \begin{cases} 1 & if \quad j \in D_3 \\ (1-\lambda) & if \quad j \in D_2 \end{cases}$$

indicateurs des variables de domaines, ($0 \leq \lambda \leq 1$)

Estimateur "Unit Multiplicity"

$$\hat{Y}_M = \sum_{j=1}^{m_{A_1}} \frac{1}{\pi_j^{A_1}} \sum_{i \in U^B} \frac{L_{ji,A_1}}{L_i^B} y_j + \sum_{j=1}^{m_{A_2}} \frac{1}{\pi_j^{A_2}} \sum_{i \in U^B} \frac{L_{ji,A_2}}{L_i^B} y_j$$

L_i^B représente le nombre total number de liens entre l'unité $j \in A_q$, (q=1,2) et l'élément i de U^B

$\pi_j^{A_q}$ représente la probabilité de sélection de l'unité j de A_q, (q=1,2)

$$L_{ji,q} = \begin{cases} 1 & \textbf{s'il y a un lien entre la jème unité de } \mathbf{A_q}, \textit{et } \textbf{l'unité i de } \mathbf{U^B} \\ \mathbf{0} & \textbf{sinon} \end{cases}$$

Estimateur “Dual Frame”

L’estimateur “Dual Frame” proposé par **Hartley (1974) peut être adapté dans le contexte du sondage indirect** :

$$\hat{Y}_H = \sum_{j \in s_{A_1}} \frac{1}{\pi_j^{A_1}} \underbrace{\frac{N_{A_1}}{\hat{N}_{A_1}} \varphi_j^{A_1}}_{C_j} y_j + \sum_{j \in s_{A_2}} \frac{1}{\pi_j^{A_2}} \underbrace{\frac{N_{A_2}}{\hat{N}_{A_2}} \varphi_j^{A_2}}_{D_j} y_j$$

$\frac{N_{A_1}}{\hat{N}_{A_1}} \quad \frac{N_{A_2}}{\hat{N}_{A_2}}$ Facteurs d’ajustement par post-stratification (ou g-poids) de chaque base de sondage

$$\varphi_j^{A_1} = \begin{cases} 1 & if \quad \delta_j^{A_2} = 0 \\ \tilde{\theta}_j^{A_1} & if \quad \delta_j^{A_2} = 1 \end{cases} \qquad \varphi_j^{A_2} = \begin{cases} 1 & if \quad \delta_j^{A_1} = 0 \\ 1 - \tilde{\theta}_j^{A_1} & if \quad \delta_j^{A_1} = 1 \end{cases}$$

$\tilde{\theta}_j^{A_1}$ j éléments de la base A_2 qui appartiennent aussi à la base A_1

$1 - \tilde{\theta}_j^{A_1}$ j éléments de la base A_1 qui appartiennent aussi à la base A_2

$\delta_j^{A_q}$ variable indiquant la base

$\pi_j^{A_q}$ représente la probabilité de sélection de l’unité j de U^{Aq} avec q=1,2.

Estimateur "Dual Frame" Hartley (1974)

$$\hat{Y}_H = \sum_{j \in s_{A_1}} \frac{1}{\pi_j^{A_1}} \underbrace{\frac{N_{A_1}}{\hat{N}_{A_1}} \varphi_j^{A_1}}_{C_j} y_j + \sum_{j \in s_{A_2}} \frac{1}{\pi_j^{A_2}} \underbrace{\frac{N_{A_2}}{\hat{N}_{A_2}} \varphi_j^{A_2}}_{D_j} y_j$$

Estimateurs "Domain Membership"

$$C_j = t_j^{A_1} z_j(\lambda) \quad \textbf{et} \quad D_j = t_j^{A_2} x_j(\lambda)$$

Estimateurs "Unit Multiplicity"

$$C_j = \sum_{i \in U^B} \frac{L_{ji,1}}{L_i^B} \quad et \quad D_j = \sum_{i \in U^B} \frac{L_{ji,2}}{L_i^B}$$

Estimateur Optimal de Deville et Lavallée (2007)

$$\hat{Y}^{opt,B} = \sum_{j=1}^{N^A} \sum_{i=1}^{N^B} \tilde{\theta}_{ji,AB}^{opt} \frac{t_j^A}{\pi_j^A} y_i$$

$$\hat{Y}^{opt,B} = \sum_{j=1}^{N^{A_1}} \sum_{i=1}^{N^B} \tilde{\theta}_{ji,A_1}^{opt} \frac{t_j^{A_1}}{\pi_j^{A_1}} y_i + \sum_{j=1}^{N^{A_2}} \sum_{i=1}^{N^B} \tilde{\theta}_{ji,A_2}^{opt} \frac{t_j^{A_2}}{\pi_j^{A_2}} y_i$$

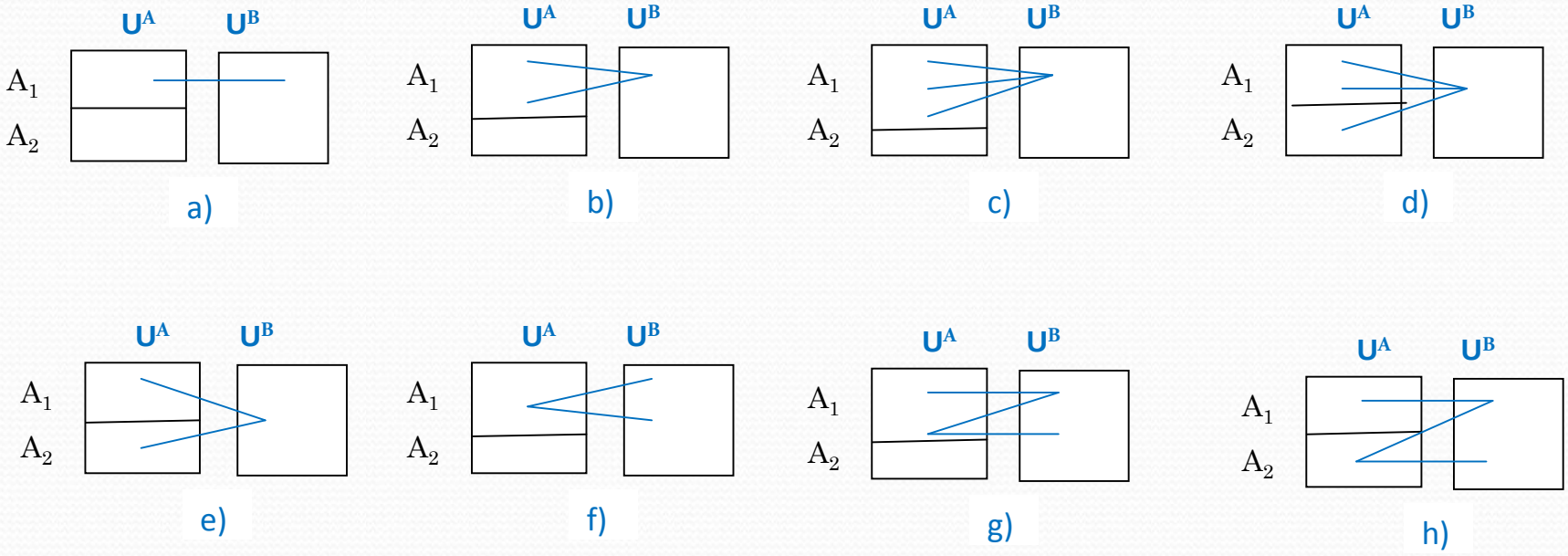
$\pi_j^{A_q}$ Probabilité de sélection de l'unité j dans U^{Aq} avec q=1,2

$\tilde{\theta}_{ji,A_q}^{opt}$ Matrice de liens optimale standardisée

$$t_j^{A_q} = \begin{cases} 1 & si\ j \in s^{A_q} \\ 0 & si\ j \notin s^{A_q} \end{cases}, q = 1,2$$

Types de liens dans le Sondage Indirect

Dans le Sondage Indirect, quand nous avons deux bases, voici les différentes combinaisons de lien qui peuvent se produire :



Comparer

Estimateur optimal de Deville et Lavallée

Estimateur "Domain Membership"

Estimateur "Unit Multiplicity"

• Dans les cas a) et f) les trois estimateurs sont les mêmes

•Dans tous les cas l'estimateur "Unit Multiplicity" et l'estimateur de Devillee et Lavallée sont les mêmes si les variances des $t_j^{A_q}$ sont égales et si les covariances sont négligeables

•Dans les cas g) et h) l'estimateur "Domain Membership" et l'estimateur de Deville et Lavallée ne peuvent être comparés car dans la théorie de Hartley nous ne pouvons avoir qu'un lien de U^A vers l'élément *i* de U^B

Variance de l'estimateur optimal de Deville et Lavallée (2007) dans le contexte de bases de sondage doubles :

$$Var\left(\hat{Y}^{opt,B}\right)=\sum_{j=1}^{N^A}\sum_{j'=1}^{N^A}\frac{\pi_{jj'}^{A_1}-\pi_j^{A_1}\pi_{j'}^{A_1}}{\pi_j^{A_1}\pi_{j'}^{A_1}}\sum_{i=1}^{N^B}\left(\tilde{\theta}_{ji,A_1}^{opt}\right)^2 y_i^2+\sum_{j=1}^{N^A}\sum_{j'=1}^{N^A}\frac{\pi_{jj'}^{A_2}-\pi_j^{A_2}\pi_{j'}^{A_2}}{\pi_j^{A_2}\pi_{j'}^{A_2}}\sum_{i=1}^{N^B}\left(\tilde{\theta}_{j'i,A_2}^{opt}\right)^2 y_i^2$$

Echantillonnage de Poisson sampling avec probabilités d'inclusion π_j^A égales
Deville and Lavallée (2006)

$$Var\left(\hat{Y}^{opt,B}\right)=\sum_{j=1}^{N^A}\sum_{j'=1}^{N^A}\frac{\pi_{jj'}^{A_1}-\pi_j^{A_1}\pi_{j'}^{A_1}}{\pi_j^{A_1}\pi_{j'}^{A_1}}y_j^2+\sum_{j=1}^{N^A}\sum_{j'=1}^{N^A}\frac{\pi_{jj'}^{A_2}-\pi_j^{A_2}\pi_{j'}^{A_2}}{\pi_j^{A_2}\pi_{j'}^{A_2}}y_j^2$$

Echantillonnage de Poisson avec probabilités d'inclusion inégales π_j^A

$$Var(\hat{Y}^{opt,B}) = \sum_{j=1}^{N^A} \frac{1-\pi_j^A}{\pi_j^A} \sum_{i=1}^{N^B} \tilde{\theta}_{ji}^{ABopt} y_i$$

où

$$\tilde{\theta}_{ji}^{ABopt} = \sum_{j=1}^{N^G} I_{jg}^{AG} \frac{\pi_g}{(1-\pi_g)\tau_i^G} I_{gi}^{GB}$$

$$\tau_i^G = \sum_{g=1}^{N_i^G} \frac{\pi_g^A}{(1-\pi_g^A)}, i = 1,2,\ldots N^B$$

N_i^G – *nombre d'unités de* U^A *liées* **à l'unité** *i of* U^B

$\frac{\pi_g^A}{(1-\pi_g^A)}$ – *correspond à une unité* j *de* U^A *qui a été préalablement liée à l'unité* g *de* U^G,

à son tour liée à l'unité i *de* U^B

$$Var(\hat{Y}^{opt,B}) = \sum_{j=1}^{N^A} \frac{1-\pi_j^A}{\pi_j^A} \left[\sum_{i=1}^{N^B} \sum_{g=1}^{N^B} I_{ji}^{AB} \frac{\pi_g^A}{(1-\pi_g^A)\,\tau_i^G} y_i \right]^2$$

$$I_{ji}^{AB} = \begin{cases} 1 & \textbf{s'il y a un lien entre la jème unité de } \mathbf{A_q}, \textit{et } \textbf{l'unité i de } \mathbf{U^B} \\ \mathbf{0} & \textbf{sinon} \end{cases}$$

Etude par Simulations

But: Comparer les variances des trois estimateurs

Données: issues de l' Eurobaromètre 68.2 (2008) relatives au Portugal

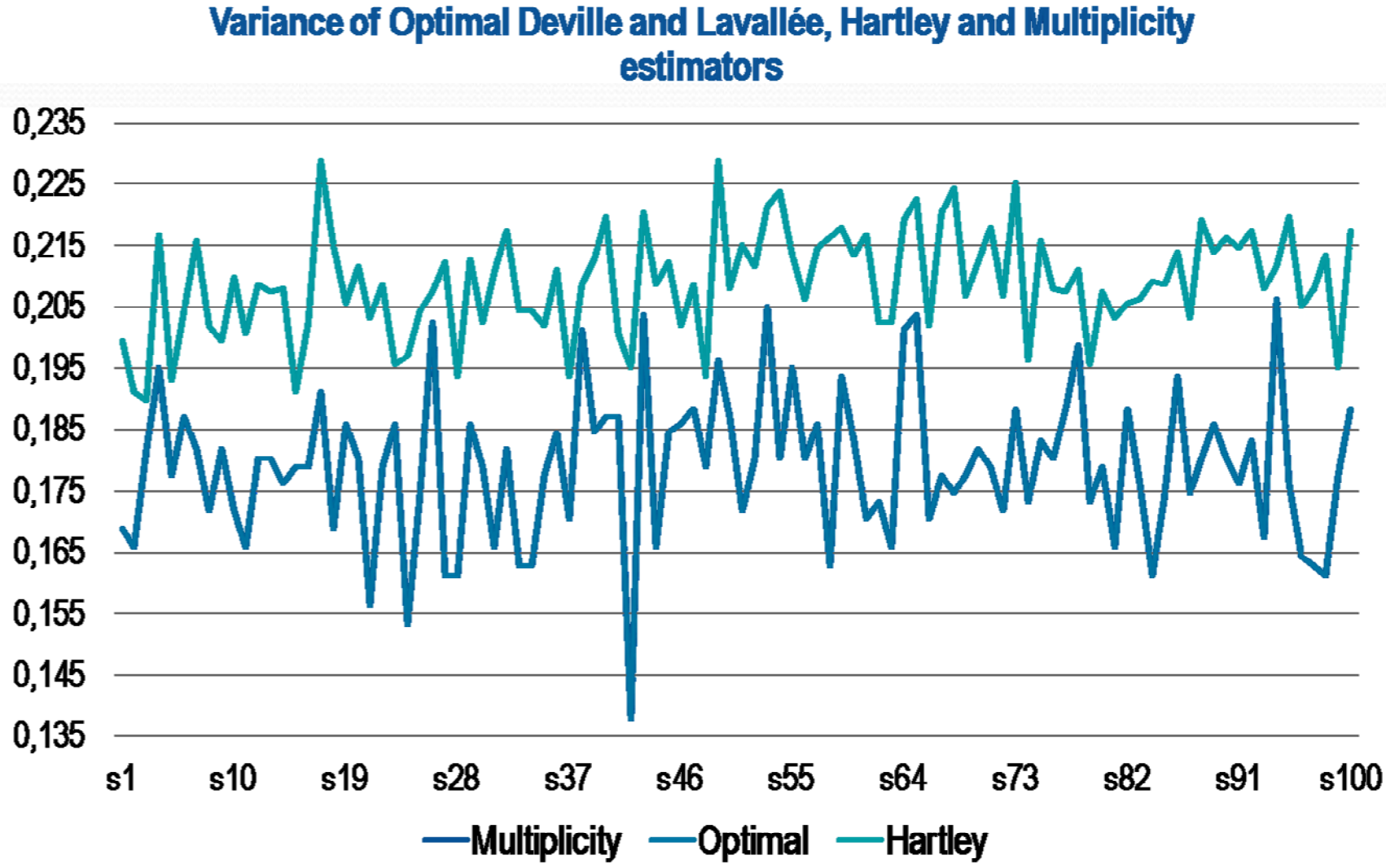
Population: 1000 ménages
Deux bases de sondage – ménages avec un téléphone fixe et un téléphone mobile

Nombre de simulations: 100 échantillons (sondage aléatoire simple)

Taille d'échantillon: 150

Paramètre d'étude: *proportion de ménages Portugais avec un accès internet à domicile*

Etude par Simulations

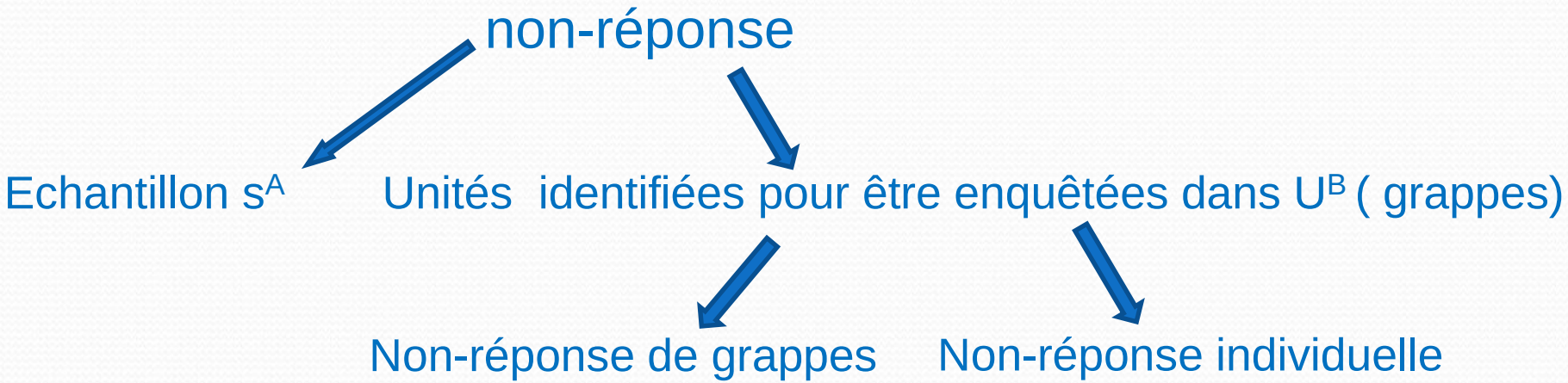


En résumé

- L'estimateur optimal et l'estimateur "Unit Multiplicity" se comportent mieux en termes de variance que l'estimateur de Hartley

- L'estimateur optimal et l'estimateur "Unit Multiplicity" ont le même comportement

Non-réponse pour le Sondage Indirect



Problèmes d'identification des liens – Impossible de déterminer si l'unité *ik* de U^B est liée à une unité *j* de U^A

Sirken et Nathan (1988) –Network Sampling

Ardilly et Le Blanc (1999;201) – GWSM weighting a survey of homeless people

Xu et Lavallée (2009) ⟹ Estimation du nombre total de liens, L_i^B, existant en utilisant des ajoustements proportionnels

Travail à venir

- Corrections proposées par Xu and Lavallée (2009) pour corriger le problème dans le contexte de bases doubles

RÉFÉRENCES

Ardilly, P., Le Blanc, P. (2001), Sampling and Weighting a Survey of Homeless Persons: A French Example, *Survey Methodology*, Vol.7, No. 1, pp. 109-118.

Bankier (1986), Estimators Based on Several Stratified Samples With Applications to Multiple Frame Surveys, *Journal of the American Statistical Association,* Vol. 81, pp.1074-1079.

Casady and Sirken (1980), A Multiplicity Estimator for Multiple Frame Sampling, *Proceedings of the Survey Reseach Methods Section, American Statistical Association,* pp. 601-605.

Deville and Lavallée (2006), Indirect sampling: The Foundations of Generalized Weight Share Method, *Survey Methodology*, Vol. 32, No.2, pp. 165-176.

European Commission (2008), *Eurobarometer 68.2 wave November 2007-January 2008*. Brussels: European Commission.

Fuller e Burmeister (1972),Estimators of samples selected from two overlapping frames, *Proceedings of the Social Statistics Section, American Statistical Association*, 245-249

Hartley, H.O. (1962), Multiple Frame Surveys, *Proceedings of the American Statistical Association, Social Statistics Section*, pp. 99-118

Hartley, H. O. (1974), Multiple Frame Surveys Methodology and Selected Applications. *Sankhyä*. C, 36, 99-118.

RÉFÉRENCES

Kalton e Anderson(1986), Sampling Rare Populations, *Journal of the Royal Statistical Society, Series A,* vol. 149, nº 1, pp. 65-82

Lavallée, P.(1995) Cross–sectional weighting of longitudinal surveys of individuals and households using weight share method. *Survey Methodology*, Vol. 21, No. 1, pp. 25-32.

Lavallée, P. (2007), *Indirect Sampling*, New York, Springer.

Lohr and Rao (2000), Inference from Dual Frame Surveys, *Journal of the American Statistical Association,* Vol. 95, nº 449, pp.271-280.

Mecatti, F. (2007), A single frame multiplicity estimator for multiple frame surveys, *Survey Methodology*, Vol. 33, No. 2, pp. 151-157

Skinner (1991), On the Efficiency of Ratio Estimation for Multiple Frame Surveys *Journal of the American Statistical Association,* Vol. 86, nº 415, pp.779-784.

Skinner and Rao (1996), Estimation on Dual Frame Surveys With Complex Designs, *Journal of the American Statistical Association,* Vol. 91, pp.349-356.

CATÓLICA
UNIVERSIDADE CATÓLICA PORTUGUESA | PORTO
Faculdade de Economia e Gestão

Merci beaucoup !!!

mmaia@porto.ucp.pt