Méthodes d'estimation sur Bases de Sondage Chevauchantes dans le cas de plans de sondage à deux degrés

Guillaume Chauvet - Guylène Tandeau de Marsac

Ensai - Insee

Journées de Méthodologie Statistique
Cité internationale universitaire de Paris
02/04/2015

ENSAI École nationale de la statistique et de l'analyse de l'information

1. Estimation pour des bases de sondage chevauchantes

2. Estimation avec un premier degré de tirage commun

3. Etude par simulations

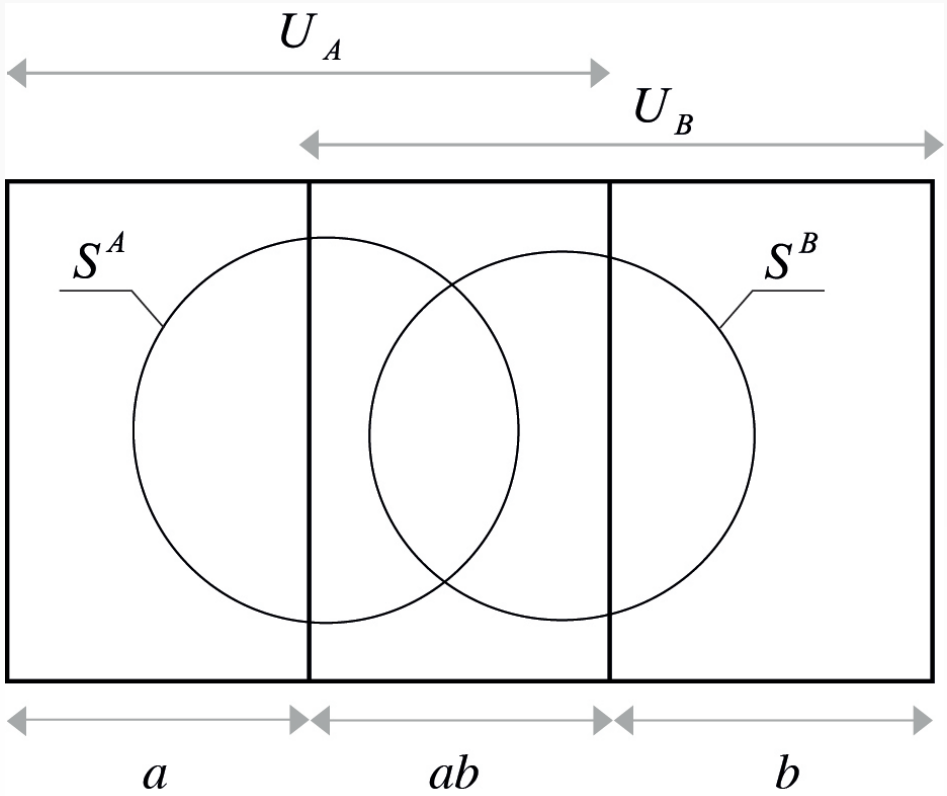
Travail issu du mémoire de Master de Statistique Publique de Guylène, qui a donné lieu à un article dans Techniques d'Enquête.

Motivé par le cas de l'Enquête Santé et Itinéraire Professionnel (2006-2010).

Estimation pour des bases de sondage chevauchantes

Principe

Population d'intérêt U, entièrement couverte par deux bases de sondage U_A (échantillon S^A) et U_B (échantillon S^B).
Objectif : combiner les deux échantillons S^A et S^B pour obtenir une bonne estimation du total $Y = \sum_{k \in U} y_k$.

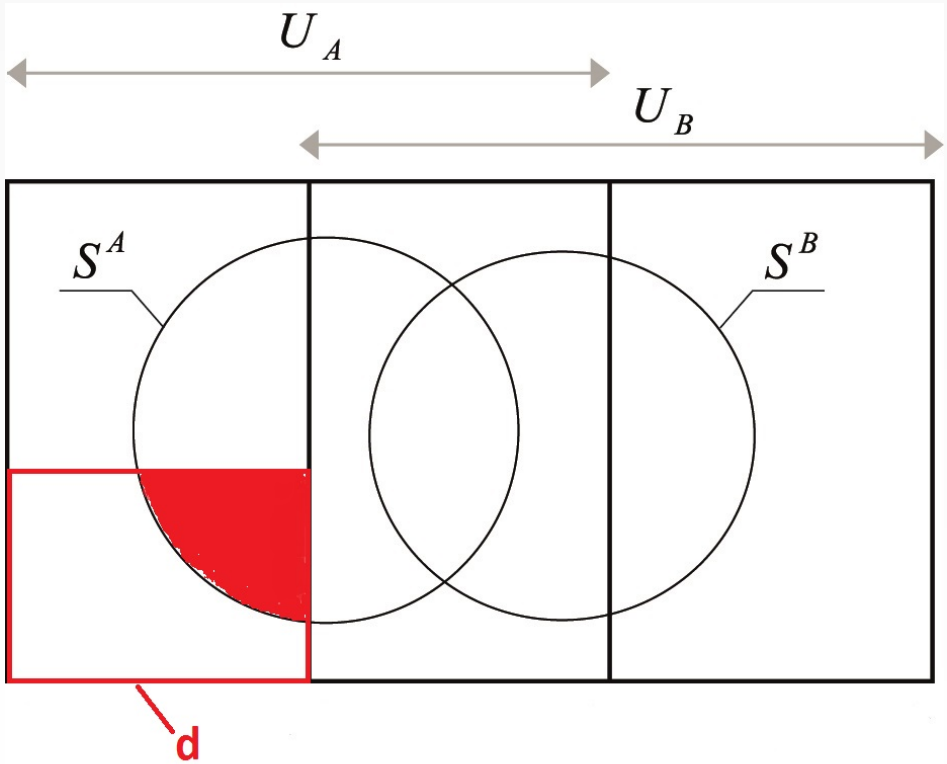


ENSAI École nationale de la statistique et de l'analyse de l'information

Notations

Si $d \subset U_A$, on estime

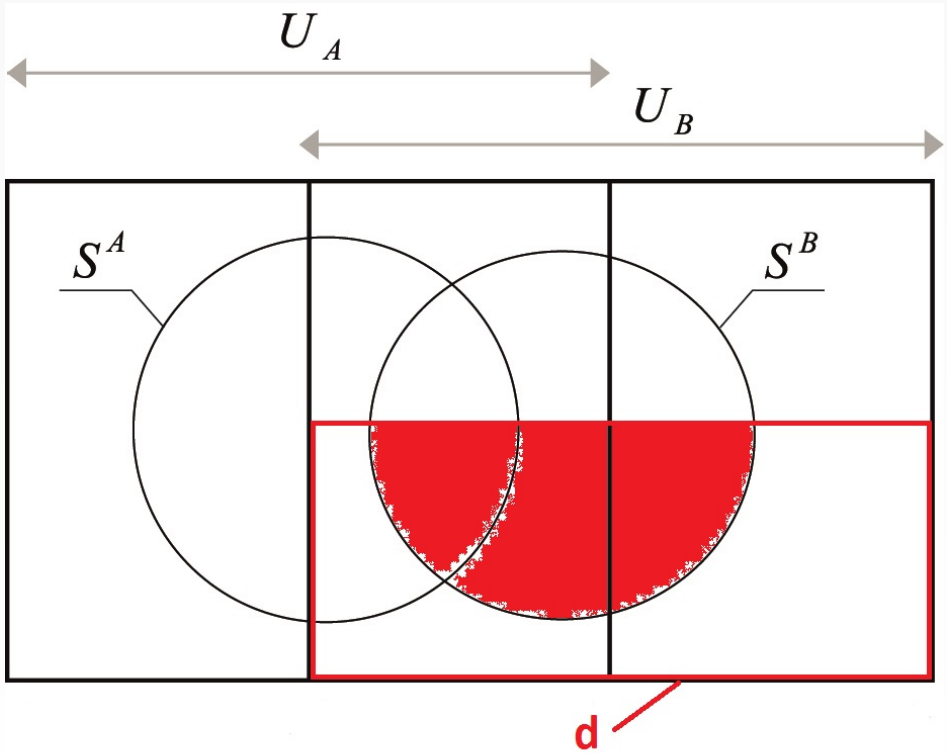
$$Y_d = \sum_{k \in d} y_k \quad \text{par} \quad \hat{Y}_d^A = \sum_{k \in S^A \cap d} d_k^A y_k \text{ basé sur } S^A.$$



Notations

Si $d \subset U_B$, on estime

$$Y_d = \sum_{k \in d} y_k \quad \text{par} \quad \hat{Y}_d^B = \sum_{k \in S^B \cap d} d_k^B y_k \text{ basé sur } S^B.$$



Estimateurs de Hartley

On réécrit

$$Y = Y_a + \theta Y_{ab} + (1-\theta) Y_{ab} + Y_b,$$
$$\Rightarrow \quad \hat{Y}_\theta = \hat{Y}_a^A + \theta \hat{Y}_{ab}^A + (1-\theta)\hat{Y}_{ab}^B + \hat{Y}_b^B.$$

Avec $\theta = 0.5$, on obtient un partage des poids couramment utilisé.

La minimisation de $V(\hat{Y}_\theta)$ conduit à

$$\theta_{opt} = \frac{V(\hat{Y}_{ab}^B) + Cov(\hat{Y}_{ab}^B, \hat{Y}_b^B) - Cov(\hat{Y}_a^A, \hat{Y}_{ab}^A)}{V(\hat{Y}_{ab}^A) + V(\hat{Y}_{ab}^B)}$$

si S^A et S^B sont indépendants. L'estimation de θ_{opt} peut être complexe.

ENSAI École nationale de la statistique et de l'analyse de l'information

Estimateur de Bankier

L'idée consiste à utiliser globalement sur l'échantillon réunion un estimateur de type Horvitz-Thompson. Les probabilités d'inclusion valent

$$\pi_k^{HT} \quad = \quad P(k \in S^A \cup S^B).$$

Si S^A et S^B sont indépendants, $\pi_k^{HT} = \pi_k^A + \pi_k^B - \pi_k^A \pi_k^B$ et on obtient l'estimateur

$$\hat{Y}_{HT} \quad = \quad \sum_{k \in S^A \cup S^B} \frac{y_k}{\pi_k^{HT}}.$$

Avantage : les poids sont les mêmes quelle que soit la variable d'intérêt.
Inconvénient : le calcul des probabilités π_k^{HT} peut être complexe.

ENSAI École nationale de la statistique et de l'analyse de l'information

Estimateur de Kalton et Anderson

L'idée consiste à utiliser globalement sur l'échantillon réunion un estimateur de type Hansen-Hurwitz. On obtient

$$\hat{Y}_{KA} = \sum_{k \in S_A} d_k^A m_k^A y_k + \sum_{k \in S^B} d_k^B m_k^B y_k$$

avec

$$m_k^A = \begin{cases} 1 & \text{si } k \in a, \\ \frac{d_k^B}{d_k^A + d_k^B} & \text{si } k \in ab, \end{cases} \quad \text{et} \quad m_k^B = \begin{cases} 1 & \text{si } k \in b, \\ \frac{d_k^A}{d_k^A + d_k^B} & \text{si } k \in ab. \end{cases}$$

Avantage : les poids sont les mêmes quelle que soit la variable d'intérêt.

ENSAI École nationale de la statistique et de l'analyse de l'information

Estimation avec un premier degré de tirage commun

Principe

Au premier degré, tirage d'un échantillon S_I d'Unités Primaires (UP), avec le poids de sondage d_{Ii} pour l'UP u_i.

Au second degré :

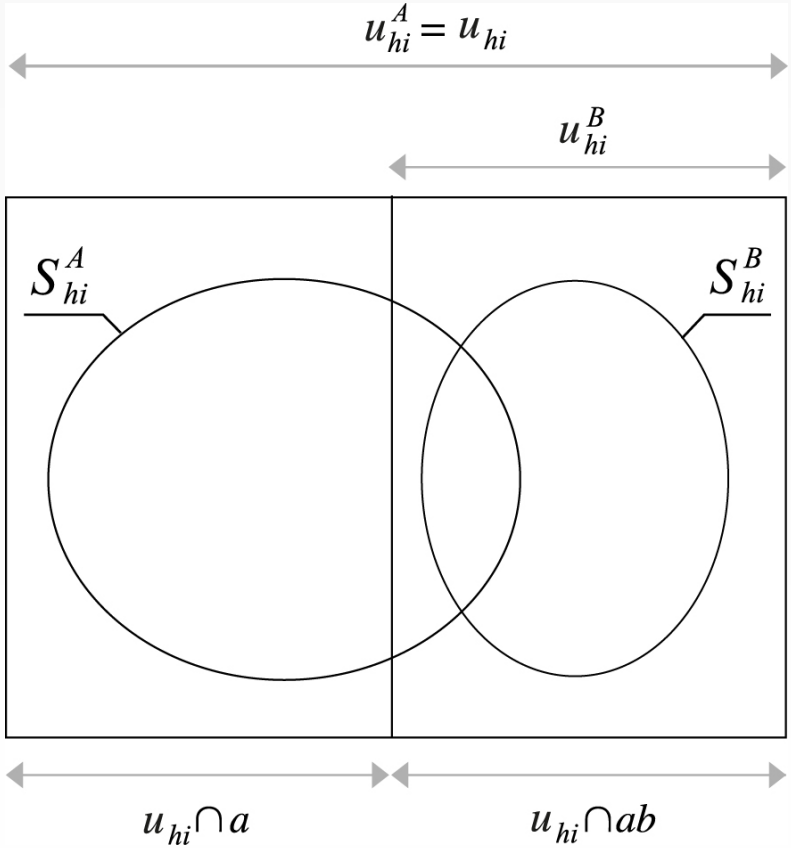
- tirage de S_i^A dans $u_i^A = u_i \cap U_A$ avec un poids de sondage $d_{k|i}^A$,
- tirage de S_i^B dans $u_i^B = u_i \cap U_B$ avec un poids de sondage $d_{k|i}^B$.

Outre les propriétés habituelles d'un plan de sondage à deux degrés, on suppose que les deux échantillons sont tirés indépendamment conditionnellement au premier degré.

ENSAI École nationale de la statistique et de l'analyse de l'information

Application

On rencontre cette configuration dans le cas d'une estimation transversale pour une enquête longitudinale, avec mise en commun des échantillons des vagues 1 et 2 pour une estimation en vague 2 (enquêtes SIP, PPV, ...).



Estimateurs de Hartley

On montre que

$$\hat{Y}_\theta = \sum_{u_i \in S_I} d_{Ii} \hat{Y}_{\theta,i} \quad \text{avec} \quad \hat{Y}_{\theta,i} \text{ estim. de Hartley pour } Y_i,$$

$$V(\hat{Y}_\theta) = \underbrace{V\left(\sum_{u_i \in S_I} d_{Ii} Y_i\right)}_{\text{indépendant de } \theta} + EV\left(\hat{Y}_\theta | S_I\right).$$

En minimisant le terme de second degré, on obtient

$$\theta_{opt} = \frac{EV(\hat{Y}^B_{ab}|S_I) + ECov(\hat{Y}^B_{ab}, \hat{Y}^B_b|S_I) - ECov(\hat{Y}^A_a, \hat{Y}^A_{ab}|S_I)}{EV(\hat{Y}^A_{ab}|S_I) + EV(\hat{Y}^B_{ab}|S_I)}.$$

ENSAI École nationale de la statistique et de l'analyse de l'information

Estimateurs de Bankier + Kalton et Anderson

L'idée consiste également à appliquer la méthode unité primaire par unité primaire. On obtient

$$\hat{Y}_{HT} = \sum_{u_i \in S_I} d_{Ii} \hat{Y}_{HT,i} \quad \text{avec} \quad \hat{Y}_{HT,i} \text{ estim. de Bankier pour } Y_i,$$

$$\hat{Y}_{KA} = \sum_{u_i \in S_I} d_{Ii} \hat{Y}_{KA,i} \quad \text{avec} \quad \hat{Y}_{KA,i} \text{ estim. de KA pour } Y_i.$$

Ce résultat est particulièrement intéressant pour la méthode optimale de Hartley, puisque l'estimateur du coefficient optimal ne nécessite que des estimateurs de variance conditionnels au premier degré.

ENSAI École nationale de la statistique et de l'analyse de l'information

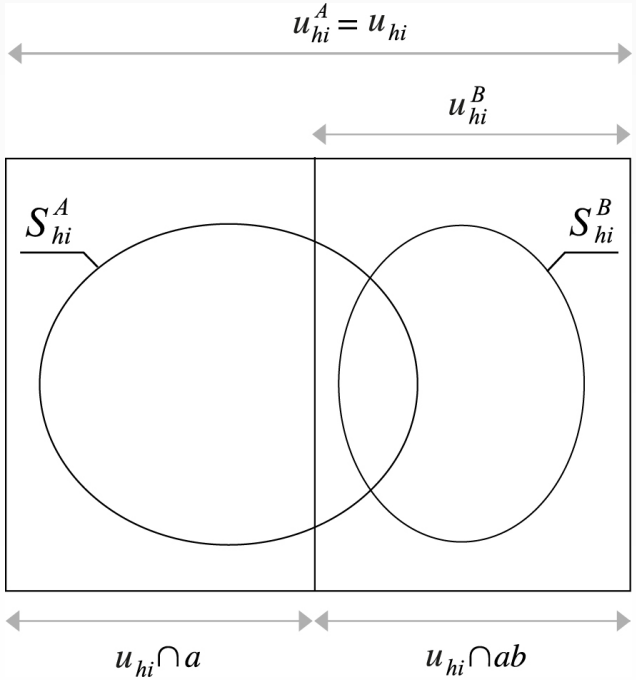
Etude par simulations

Population simulée (configuration proche de SIP)

Nous générons une population de $M = 200$ UP regroupées en $H = 4$ strates U_{Ih} de même taille. Dans chacune, nous générons pour chaque US $k \in u_{hi}$:

$$y_k \quad = \quad \mu_h + \sigma_h \; v_{hi} + \{\rho^{-1}(1-\rho)\}^{0.5} \sigma_h \; v_k, \qquad (1)$$

où $v_{hi}, v_k \sim \mathcal{N}(0,1)$. Le paramètre ρ est choisi de façon à obtenir un coefficient de corrélation intra-grappes approximativement égal à 0.2.



ENSAI École nationale de la statistique et de l'analyse de l'information

Plan de sondage et paramètres d'intérêt

Dans chaque population, nous sélectionnons $B = 10,000$ échantillons selon un plan à deux degrés, avec :

- un sondage aléatoire simple stratifié au 1er degré ($m_h = 5$ ou $m_h = 25$),
- un sondage aléatoire simple au 2nd degré ($n_{hi}^A = 10$ ou $n_{hi}^A = 40 + n_{hi}^B = 5$ ou $n_{hi}^B = 20$).

On compare l'EQM des estimateurs :

- de Hartley avec $\theta = 0.5$ (HART1),
- optimal (estimé) de Hartley (HART2),
- de Kalton et Anderson (KALT),
- de Bankier (BANK),
- utilisant S^A (HTA).

ENSAI École nationale de la statistique et de l'analyse de l'information

Résultats obtenus pour l'EQM ($\times 10^9$)

m_h	n_{hi}^A	n_{hi}^B	HART1	HART2	KALT	BANK	HTA
5	10	5	7.76	5.70	7.79	8.56	5.75
5	10	20	7.57	5.57	11.36	12.75	5.75
5	40	5	5.01	4.51	4.57	4.81	4.52
5	40	20	4.65	4.33	4.66	5.22	4.52
25	10	5	1.19	0.78	1.20	1.34	0.78
25	10	20	1.17	0.78	1.94	2.22	0.78
25	40	5	0.62	0.51	0.52	0.57	0.51
25	40	20	0.58	0.51	0.58	0.68	0.51

Résultats obtenus pour l'EQM ($\times 10^9$)

m_h	n_{hi}^A	n_{hi}^B		HART2			HTA
5	10	5		5.70			5.75
5	10	20		5.57			5.75
5	40	5		4.51			4.52
5	40	20		4.33			4.52
25	10	5		0.78			0.78
25	10	20		0.78			0.78
25	40	5		0.51			0.51
25	40	20		0.51			0.51

Résultats obtenus pour l'EQM ($\times 10^9$)

m_h	n_{hi}^A	n_{hi}^B	HART1		KALT	BANK	
5	10	5	7.76		7.79	8.56	
5	10	20	7.57		11.36	12.75	
5	40	5	5.01		4.57	4.81	
5	40	20	4.65		4.66	5.22	
25	10	5	1.19		1.20	1.34	
25	10	20	1.17		1.94	2.22	
25	40	5	0.62		0.52	0.57	
25	40	20	0.58		0.58	0.68	

Résultats obtenus pour l'EQM ($\times 10^9$)

m_h	n_{hi}^A	n_{hi}^B			KALT	BANK	
5	10	5			7.79	8.56	
5	10	20			11.36	12.75	
5	40	5			4.57	4.81	
5	40	20			4.66	5.22	
25	10	5			1.20	1.34	
25	10	20			1.94	2.22	
25	40	5			0.52	0.57	
25	40	20			0.58	0.68	

Références

- Bankier, M.D. (1986). Estimators Based on Several Stratified Samples With Applications to Multiple Frame Surveys. JASA, 81, p. 1074-1079.
- Chauvet G., Tandeau de Marsac, G. (2014). Méthodes d'estimation sur bases de sondage multiples dans le cadre de plans de sondage à deux degrés. Techniques d'enquête, 40, 2, p. 367-378.
- Hartley, H.O. (1962). Multiple frame surveys. Proceedings of the Social Statistics Section, American Statistical Association, p. 203-206.
- Kalton, G., et Anderson, D.W. (1986). Sampling Rare Populations. JRSS A, 149, p. 65-82.
- Lavallée, P. (2002). Le Sondage Indirect, ou la Méthode généralisée du partage des poids. Éditions de l'Université de Bruxelles (Belgique) et Éditions Ellipses (France).
- Lohr, S.L. (2009). Multiple frame surveys. In Handbook of Statistics, Sample Surveys : Design, Methods and Applications, (Eds., D. Pfeffermann and C.R. Rao), Vol. 29A, p. 71-88.

ENSAI École nationale de la statistique et de l'analyse de l'information