Estimateur composite du total de la population sous l'approche basée sur le plan et le modèle

Mohammed El Haj Tirari

INSEA, Rabat-Instituts, B.P. 6217, Rabat, Maroc,
mtirari@hotmail.fr

Résumé. L'estimation du total de la population est parmi les problématiques ayant suscitée beaucoup d'intérêt en théorie des sondages. En effet, le fait de bien estimer un total conduit à l'amélioration de l'estimation de tout paramètre de la population qui peut s'écrire comme une fonction de totaux. Plusieurs estimateurs ont été proposés pour estimer un total dont l'estimateur d'Horvitz-Thompson (1952) et celui par calage (Deville, 1992) sont parmi les plus utilisés en pratique. Cependant, chacun de ces deux estimateurs possède des avantages et des inconvénients. Nous proposons un nouvel estimateur composite des estimateurs HT et par calage dont l'élaboration est faite de telle sorte à tenir compte des situations les plus favorables à l'utilisation de chacun de ces deux estimateurs.

Mots-clés. Calage, estimateur HT, modèle de superpopulation, Information auxiliaire, aproche modèle.

1 Introduction

Pour estimer le total d'une population, plusieurs estimateurs ont été proposés. Le choix entre ces estimateurs dépend principalement de la disponibilité des variables auxiliaires, le type de ces variables et le degré de leur lien avec la variable d'intérêt. Parmi ces estimateurs, on fait souvent recours aux estimateurs HT et par calage. En effet, l'estimateur HT est utilisé quand on ne dispose pas d'information auxiliaire et il a l'avantage d'être le seul estimateur sans biais sous le plan de sondage. En présence d'information auxiliaire, l'estimateur par calage est le plus utilisé en pratique pour en tenir compte dans le but d'améliorer la précision des estimations produites. Cependant, cet estimateur peut ne pas convenir à toutes les variables d'intérêt de l'enquête, en particulier celles qui ne sont pas liées aux variables auxiliaires utilisées dans le calage et il peut donc aboutir à des estimations du total moins précises que celles obtenues avec l'estimateur HT. En considérant l'approche basée sur le plan et le modèle, nous proposons dans ce travail un estimateur composite des estimateurs HT et par calage de telle sorte à tirer profit des situations favorables à l'utilisation de chacun des deux estimateurs.

Pour l'élaboration de cet estimateur composite du total de la population, on sait placer dans le cas où on dispose d'un ensemble de variables auxiliaires tout en supposant que

le lien entre ces dernières et la variable d'intérêt peut être représenté par un modèle de superpopulation telle que un modèle de régression linéaire. Ainsi, pour l'estimateur composite proposé, la détermination du coefficient permettant de combiner les estimateurs HT et par calage est basée sur les EQM condionnellement au plan et au modèle de ces deux estimateurs.

2 Notations et définitions

Soit $U = \{1, \ldots, N\}$ une population de taille N à partir de laquelle on sélectionne un échantillon s de taille n selon un plan de sondage $p(\cdot)$ dont les probabilités d'inclusion d'ordre un et deux sont données respectivement par π_k et π_{kl}. On s'intéresse à une variable d'intérêt $\mathbf{y} = (y_1, \ldots, y_N)^{'}$ en ayant comme objectif l'estimation de son total :

$$t_y = \sum_{k \in U} y_k$$

On dispose de p variables auxilaires $X_1, \ldots, X_p$ dont les valeurs peuvent être représentées pour tout $k \in U$ par les vecteurs $\mathbf{x}_k = \left(x_{k1}, \ldots, x_{kp}\right)^{'}$. On suppose que le vecteur des totaux

$$t_{\mathbf{x}} = \sum_{k \in U} \mathbf{x}_k$$

des p variables auxiliaires $(X_1, \ldots, X_p)$ est connu.

Sous l'approche basée sur le modèle, on suppose que les valeurs de la variable d'intérêt $\mathbf{y}$ sont les réalisations d'un modèle de superpopulation ξ définit par

$$y_k = \mathbf{x}_k^{'}\boldsymbol{\beta} + \epsilon_k$$

avec

$$\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{'},$$
$$E_\xi(\epsilon_k) = 0, \;\; var_\xi(\epsilon_k) = \sigma_k^2 \;\text{ et }\; cov_\xi(\epsilon_k, \epsilon_l) = 0.$$

E_ξ, var_ξ et cov_ξ représentent respectivement l'espérence, la variance et la covariance sous le modèle.

Pour estimer le total t_y d'une variable d'intérêt $\mathbf{y}$, on considère la classe des estimateurs linéaires qui peuvent s'écrire

$$\widehat{t}_{yw} = \sum_{k \in S} w_{kS} y_k$$

où w_{kS} sont des poids qui peuvent dépendre de l'échantillon.

Dans ce qui suit, nous allons considérer l'approche basée sur le plan et le modèle. Sous cette approche, la précision d'un estimateur linéaire est mesurée en considérant l'Ecart Quadratique Moyen anticipé définit par

$$EQM_{p\xi}(\widehat{t}_{yw}) = E_p E_\xi(\widehat{t}_{yw} - t_y)^2$$

Ainsi, on peut montrer que l'EQM anticipé de l'estimateur linéaire $\widehat{t}_{yw}$ est donné par (Nedyalkova et Tillé, 2008) :

$$EQM_{p\xi}(\widehat{t}_{yw}) = E_p\left(\sum_{k\in S} w_{kS}\mathbf{x}_k'\boldsymbol{\beta} - \sum_{k\in U}\mathbf{x}_k'\boldsymbol{\beta}\right)^2 + \sum_{k\in U}\sigma_k^2\left[var_p(w_{kS}I_k) + (R_{kS}-1)^2\right]$$

où $I_k = 1$ pour $k \in S$ et 0 sinon.

$$R_{kS} = E_p(w_{kS}I_k) = \frac{E_p(w_{kS}I_k \mid I_k = 1)}{d_k}$$

avec $d_k = \frac{1}{\pi_k}$. On note que R_{kS} est égale à 1 quand l'estimateur linéaire est sans biais sous le plan.

3 L'EQM anticipé des estimateurs par calage et d'HT

Un estimateur linéaire est dit calé sur les variables auxiliaires $X_1, \ldots, X_p$ si et seulement si ses poids w_{kS} vérifient

$$\sum_{k\in S} w_{kS}\mathbf{x}_k = \sum_{k\in U}\mathbf{x}_k$$

On note que les équations de calage rendent l'estimateur par calage $\widehat{t}_{yC}$ sans biais sous le modèle :

$$E_\xi\left(\widehat{t}_{yC} - t_y\right) = \sum_{k\in S} w_{kS}\mathbf{x}_k'\boldsymbol{\beta} - \sum_{k\in U}\mathbf{x}_k'\boldsymbol{\beta} = 0$$

Ainsi, l'EQM anticipé de l'estimateur par calage est donné par :

$$EQM_{p\xi}(\widehat{t}_{yC}) = \sum_{k\in U}\sigma_k^2\left[\frac{V_{kS}}{d_k} + R_{kS}^2(d_k - 1) + (R_{kS}-1)^2\right] \quad (1)$$

où $V_{kS} = var_p(w_{kS} \mid I_k = 1)$. Notons que l'expression (1) de $EQM_{p\xi}(\widehat{t}_{yC})$ permet de mettre en évidence les deux critères dont dépend la précision de l'estimateur par calage $\widehat{t}_{yC}$. Le premier est celui correspondant au modèle de Superpopulation ξ à travers sa variance résiduelle qui diminue quand la variable d'intérêt et les variables de calage sont corrélées entre elles (réduction de la variance de $\widehat{t}_{yC}$). Le second critère est représenté par les rapports de poids R_{kS} qui deviennent importants quand les poids de calage sont très différents de ceux de sondage (augmentation du biais de $\widehat{t}_{yC}$).

En ce qui concerne l'estimateur d'HT, son EQM anticipé est donné par :

$$EQM_{p\xi}(\widehat{t}_{y\pi}) = \sum_{k\in U}\sum_{l\in U}(\pi_{kl} - \pi_k\pi_l)d_k\mathbf{x}_k'\boldsymbol{\beta}d_l\mathbf{x}_l'\boldsymbol{\beta} + \sum_{k\in U}\sigma_k^2 d_k(1-\pi_k) \quad (2)$$

Comme l'expression (2) de l'EQM anticipé de $\widehat{t}_{y\pi}$ dépend des probabilités π_{kl} qui sont généralement inconnues et difficiles à calculer, une approximation de (2) peut être donnée par

$$\widetilde{EQM}_{p\xi}(\widehat{t}_{y\pi}) = \sum_{k\in U} c_k (d_k \mathbf{x}_k' \boldsymbol{\beta})^2 - \frac{1}{h}\left(\sum_{k\in U} c_k d_k \mathbf{x}_k' \boldsymbol{\beta}\right)^2 + \sum_{k\in U} \sigma_k^2 d_k (1-\pi_k)$$

où $h = \sum_{k\in U} c_k$ avec $c_k = N\pi_k(1-\pi_k)/(N-1)$.

4 Estimateur composite du total de la population

Afin de pouvoir tenir compte des avantages et inconvénients des deux estimateurs par calage ($\widehat{t}_{yC}$) et d'HT ($\widehat{t}_{y\pi}$), nous proposons d'estimer le total de la population en utilisant l'estimateur composite suivant :

$$\widehat{t}_{yComp} = \widehat{\theta}\widehat{t}_{yC} + (1-\widehat{\theta})\widehat{t}_{y\pi}$$

où

$$\widehat{\theta} = \frac{\widehat{\widetilde{EQM}}_{p\xi}(\widehat{t}_{yC})}{\widehat{\widetilde{EQM}}_{p\xi}(\widehat{t}_{y\pi})}$$

avec

$$\widehat{\widetilde{EQM}}_{p\xi}(\widehat{t}_{yC}) = \sum_{k\in s} d_k \widehat{\sigma}_k^2 \left[\frac{(w_{kS}-d_k)^2}{d_k} + \widehat{R}_{kS}^2(d_k-1) + (\widehat{R}_{kS}-1)^2\right]$$

$\widehat{R}_{kS} = (w_{kS}/d_k$ et

$$\widehat{\widetilde{EQM}}_{p\xi}(\widehat{t}_{y\pi}) = \sum_{k\in s} \widetilde{c}_k (d_k \mathbf{x}_k' \boldsymbol{\beta})^2 - \frac{1}{\widetilde{h}}\left(\sum_{k\in s} \widetilde{c}_k d_k \mathbf{x}_k' \boldsymbol{\beta}\right)^2 + \sum_{k\in s} \widehat{\sigma}_k^2 d_k (d_k-1)$$

$\widetilde{c}_k = n(1\pi_k)/(n-1)$ et $\widetilde{h} = \sum_{k\in s} \widetilde{c}_k$.

Ainsi, l'estimateur composite $\widehat{t}_{yComp}$ proposé permet de mettre à profit le gain en précision obtenu en faisant recours au calage tout en restant robuste lorsque la perte en terme du biais introduit dépasse la gain réalisé à travers la réduction de la variance. Une étude des propriétés de cet estimateur sera présenté dans cette communication.

Bibliographie

[1] Deville, J.-C. et Särndal, C.-E. (1992), Calibration estimators in survey sampling, Journal of the American Statistical Association, 87(418), 376-382.

[2] El Haj Tirari, M. (2012). Critère du choix des variables auxiliaires à utiliser dans l'estimateur par calage. Septième Colloque Francophone sur les sondages, Rennes, France.
[3] El Haj Tirari, M. (2016). Doit-on utiliser toujours la pondération de calage ? Neuvième Colloque Francophone sur les sondages, Gatineau, Canada.
[3] Henry, K. A. et Skinner, R. (2015), A design effect measure for calibration weighting in single-stage samples, Survey Methodology, 41, N 2, 315-331.
[4] Horvitz, D., et Thompson, D. (1952). A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association, 47, 663-685.
[5] Kott, P. (2009). Calibration weighting : Combining probability samples and linear prediction models. Dans *Handbook of Statistics, Sample Surveys : Design, Methods and Application*, (éds., D. Pfeffermann et C.R. Rao), 29B, Amsterdam : Elsevier BV.