Aust. N. Z. J. Stat. 2014 doi: 10.1111/j.1467-842X.XXX

Estimation of the population total in Dual Frames Survey

YUICHI HIROSE, RICHARD ARNOLD AND DANIEL FERNANDEZ

Victoria University of Wellington

Summary

xxxxxxxxxxxxxxxx

Key words: xxxxxxx

1. Introduction

Many authors (Hartley H.O. (1962, 1974), Kalton & Anderson (1986), Bankier (1986), Fuller & Burmeister (1972), Skinner (1991), Skinner & Rao (1996), Lohr & Rao (2006), Metcalf & Scott (2009)) have proposed method for estimating population totals and means for dual or multiple frame surveys. See Lohr & Rao (2000) and Lohr (2007, 2013) for overview of the various methods for dual and multiple frame surveys.

This paper aim is also to estimate the population total from dual frame surveys. We propose a method of estimating weights used in Metcalf & Scott (2009) that minimize the variance of the estimator of the population total. In simulation studies we demonstrate that the proposed method is efficient among the existing method of estimating the population total. The focus is on dual frame surveys but the method is applicable to multi frame surveys.

2. Notations

For the setting of the problem and notations, we follow Hartley H.O. (1962, 1974), Lohr & Rao (2000, 2006) and Metcalf & Scott (2009).

Suppose we have two sampling frames, A and B, together cover the population of interest $\mathcal{U}$; i.e., $\mathcal{U} = A \cup B$. Two independent samples are taken from frames A and B, and two samples are combined to for the inference of interest.

The population can be decide into three mutually exclusive components:

$$\mathcal{U} = A \cup B = a \cup ab \cup b,$$

where $a = A \setminus B$, $b = B \setminus A$ and $ab = A \cap B$. Let N be the number of units in the population $\mathcal{U}$.

Two independent samples are taken from A and B respectively. The sample from frame A is denoted by s^A with the sample size n^A and the sample from B is denoted by s^B with the sample size n^B. For each unit $i \in s^A \cup s^B$ in the samples, we observe y_i, a variable of interest.

xxxxxx, xxxxxxx.

Acknowledgment. This class file was developed by Sunrise Setting Ltd, Paignton, Devon, UK. Website: www.sunrise-setting.co.uk

†Please ensure that you use the most up to date class file, available from the ANZS Home Page at http://onlinelibrary.wiley.com/journal/10.1111/(ISSN)1467-842X

For all $i \in \mathcal{U}$, let $I_i^A = I(i \in s^A)$ and $I_i^B = I(i \in s^B)$. Then the inclusion probabilities are defined by

$$\pi_i^A = P(i \in s^A) = E(I_i^A), \;\; \pi_i^B = P(i \in s^B) = E(I_i^B). \tag{1}$$

Also define the joint inclusion probabilities

$$\pi_{ij}^A = P(i,j \in s^A) = E(I_i^A I_j^A), \;\; \pi_{ij}^B = P(i,j \in s^B) = E(I_i^B I_j^B). \tag{2}$$

We assume

$$\left\{ \begin{array}{ll} \pi_i^A > 0 & \text{if } i \in A = a \cup ab \\ \pi_i^A = 0 & \text{if } i \in b \end{array} \right. , \quad \left\{ \begin{array}{ll} \pi_i^B > 0 & \text{if } i \in B = b \cup ab \\ \pi_i^B = 0 & \text{if } i \in a \end{array} \right. , \tag{3}$$

and define weights:

$$w_i^A = \left\{ \begin{array}{ll} \frac{1}{\pi_i^A} & \text{if } i \in A = a \cup ab \\ 0 & \text{if } i \in b \end{array} \right. , \quad w_i^B = \left\{ \begin{array}{ll} \frac{1}{\pi_i^B} & \text{if } i \in B = b \cup ab \\ 0 & \text{if } i \in a \end{array} \right. . \tag{4}$$

Note that since s^A and s^B are two independent samples,

$$\text{Cov}(I_i^A, I_j^B) = 0, \quad \text{for all } i,j \in \mathcal{U}.$$

3. Estimation of the population total

We use the weights proposed in Metcalf & Scott (2009). In the method, they introduced a new parameter θ_i for each unit $i \in \mathcal{U} = a \cup ab \cup b$ in the population, which is given by

$$\theta_i = \left\{ \begin{array}{ll} 1 & \text{if } i \in a \\ \text{a constant between 0 and 1 } (0 \le \theta_i \le 1) & \text{if } i \in ab \\ 0 & \text{if } i \in b. \end{array} \right. \tag{5}$$

Then adjusted weights are given by

$$\tilde{w}_i(\theta_i) = \left\{ \begin{array}{ll} \theta_i w_i^A & \text{if } i \in s^A \\ (1-\theta_i) w_i^B & \text{if } i \in s^B. \end{array} \right. \tag{6}$$

Their proposed estimator of the population total is

$$\tilde{Y}(\theta) = \sum_{i \in s^A \cup s^B} \tilde{w}_i(\theta_i) y_i,$$

and it is shown to be unbiased for any choice of $\theta = \{\theta_i : i \in s^A \cup s^B\}$ satisfying (5): Using (4), (5) and (6), the expected value with respect to the inclusion probabilities $\pi_i^A = E(I_i^A)$, $\pi_i^B = E(I_i^B)$ is

$$\begin{aligned} E(\tilde{Y}(\theta)) &= E\left(\sum_{i \in \mathcal{U}} \{ I_i^A \theta_i w_i^A + I_i^B (1-\theta_i) w_i^B \} y_i \right) \\ &= \sum_{i \in \mathcal{U}} \{ \pi_i^A \theta_i w_i^A + \pi_i^B (1-\theta_i) w_i^B \} y_i = \sum_{i \in \mathcal{U}} Y_i. \end{aligned}$$

3.1. Estimation of the variance with θ

We estimate θ as the minimizer of the variance

$$\mathbf{V}_{\tilde{Y}(\theta)} = \mathrm{Var}\left(\sum_{i\in s^A\cup s^B}\tilde{w}_i(\theta_i)y_i\right),$$

where the variance is taken with respect to the inclusion probabilities π_i^A, π_i^B.

Let $\alpha_i^A = w_i^A y_i$ and $\alpha_i^B = w_i^B y_i$. Then by definitions (1) and (6), the variance $\mathbf{V}_{\tilde{Y}(\theta)}$ is equal to

$$\begin{aligned}\mathbf{V}_{\tilde{Y}(\theta)} &= \mathrm{Var}\left(\sum_{i\in\mathcal{U}}\{I_i^A\theta_i w_i^A + I_i^B(1-\theta_i)w_i^B\}y_i\right)\\ &= \mathrm{Var}\left(\sum_{i\in\mathcal{U}}\{\theta_i\alpha_i^A I_i^A + (1-\theta_i)\alpha_i^B I_i^B\}\right)\\ &= \sum_{i,j\in\mathcal{U}}\{\theta_i\theta_j\alpha_i^A\alpha_j^A\mathrm{Cov}(I_i^A,I_j^A) + (1-\theta_i)(1-\theta_j)\alpha_i^B\alpha_j^B\mathrm{Cov}(I_i^B,I_j^B)\},\end{aligned} \tag{7}$$

where we used $\mathrm{Cov}(I_i^A, I_j^B) = 0$ for all i, j.

The variance $\mathbf{V}_{\tilde{Y}(\theta)}$ is estimated by

$$\begin{aligned}\hat{\mathbf{V}}_{\tilde{Y}(\theta)} &= \sum_{i,j\in s^A}(\pi_{ij}^A)^{-1}\theta_i\theta_j\alpha_i^A\alpha_j^A\mathrm{Cov}(I_i^A,I_j^A)\\ &\quad+\sum_{i,j\in s^B}(\pi_{ij}^B)^{-1}(1-\theta_i)(1-\theta_j)\alpha_i^B\alpha_j^B\mathrm{Cov}(I_i^B,I_j^B).\end{aligned}$$

Determining θ:

The derivative of $\frac{N^2}{2}\times$ (7) with respect to θ_i, $i \in ab$, is

$$\begin{aligned}\frac{N^2}{2}\frac{\partial}{\partial\theta_i}\mathbf{V}_{\tilde{Y}(\theta)} &= \sum_{j\in\mathcal{U}}\{\theta_j\alpha_i^A\alpha_j^A\mathrm{Cov}(I_i^A,I_j^A) - (1-\theta_j)\alpha_i^B\alpha_j^B\mathrm{Cov}(I_i^B,I_j^B)\}\\ &= \theta_i(\alpha_i^A)^2\mathrm{Var}(I_i^A) - (1-\theta_i)(\alpha_i^B)^2\mathrm{Var}(I_i^B) + C_i(\theta),\end{aligned}$$

where

$$C_i(\theta) = \sum_{j\in\mathcal{U},j\neq i}\{\theta_j\alpha_i^A\alpha_j^A\mathrm{Cov}(I_i^A,I_j^A) - (1-\theta_j)\alpha_i^B\alpha_j^B\mathrm{Cov}(I_i^B,I_j^B)\}$$

Solution to equation $\frac{N^2}{2}\frac{\partial}{\partial\theta_i}\mathbf{V}_{U_\theta} = 0$ is

$$\hat{\theta}_i = \frac{(\alpha_i^B)^2\mathrm{Var}(I_i^B) - \hat{C}_i(\theta)}{(\alpha_i^A)^2\mathrm{Var}(I_i^A) + (\alpha_i^B)^2\mathrm{Var}(I_i^B)},$$

where $\hat{C}_i(\theta)$ is an estimator of $C_i(\theta)$ which is given by

$$\begin{aligned}\hat{C}_i(\theta) &= \sum_{j\in s^A}w_j^A\theta_j\alpha_i^A\alpha_j^A\mathrm{Cov}(I_i^A,I_j^A) - \sum_{j\in s^B}w_j^B(1-\theta_j)\alpha_i^B\alpha_j^B\mathrm{Cov}(I_i^B,I_j^B)\\ &\quad-\theta_i(\alpha_i^A)^2\mathrm{Var}(I_i^A) + (1-\theta_i)(\alpha_i^B)^2\mathrm{Var}(I_i^B).\end{aligned}$$

Recall $\mathrm{Var}(I_i^A) = \pi_i^A(1-\pi_i^A)$, $\mathrm{Var}(I_i^B) = \pi_i^B(1-\pi_i^B)$, $\mathrm{Cov}(I_i^A, I_j^A) = \pi_{ij}^A - \pi_i^A\pi_j^A$ and $\mathrm{Cov}(I_i^B, I_j^B) = \pi_{ij}^B - \pi_i^B\pi_j^B$. Then, using (4),

$$\hat{\theta}_i = \frac{w_i^B(1-\pi_i^B) - \hat{C}_i^*(\theta)}{w_i^A(1-\pi_i^A) + w_i^B(1-\pi_i^B)} \tag{8}$$

where

$$\begin{aligned}
\hat{C}_i^*(\theta) &= \frac{\hat{C}_i(\theta)}{y_i^2} \\
&= \sum_{j\in s^A} \theta_j w_i^A (w_j^A)^2 (\pi_{ij}^A - \pi_i^A\pi_j^A)\frac{y_j}{y_i} \\
&\quad - \sum_{j\in s^B} (1-\theta_j) w_i^B (w_j^B)^2 (\pi_{ij}^B - \pi_i^B\pi_j^B)\frac{y_j}{y_i} \\
&\quad -\theta_i w_i^A(1-\pi_i^A) + (1-\theta_i) w_i^B(1-\pi_i^B).
\end{aligned}$$

Therefore under the constraint (5), for $i \in a$, $\hat{\theta}_i = 1$, for $i \in b$, $\hat{\theta}_i = 0$ and for $i \in ab$, $\hat{\theta}_i$ is given by (8).

An explicit approximate solution can be obtained by putting $\mathrm{Cov}(I_i^A, I_j^A) = \mathrm{Cov}(I_i^B, I_j^B) = 0$ for all $i \neq j$ and $\mathrm{Cov}(I_i^A, I_j^B) = 0$ for all i, j. In this case $C_i(\theta) = 0$ and (8) is reduced to

$$\hat{\theta}_i = \frac{w_i^B(1-\pi_i^B)}{w_i^A(1-\pi_i^A) + w_i^B(1-\pi_i^B)}. \tag{9}$$

4. Example

5. Discussion

References

BANKIER M.D. (1986) *Estimators based on several stratified samples with applications to multiple frame surveys*, J. R, Stat. Soc, Ser. A **149**, 65–82.

FULLER, W. (2009) *Sampling Statistics*, New York, John Wiley and Sons.

FULLER, W.A. & BURMEISTER, L.F. (1972) *Estimates from samples selected from two overlapping frames*, Proceedings of the Social Statistics Section, American Statistical Association, Washington, DC, 245–249.

HARTLEY, H.O. (1962) *Multiple frame surveys*, Proceedings of the Social Statistics Section, American Statistical Association, Washington, DC, 203–206.

HARTLEY, H.O. (1974) *Multiple frame methodology and selected applications*, Sankhya, Series C, **36**, 99–118.

KALTON, G.W. & ANDERSON, D.W. (1986) *Sampling rare populations*, J. Amer. Statist. Assoc. **94**, 271–280.

LOHR, S.L. (2007) *Recent developments in multiple frame surveys*, Proceedings of the Survey Research Methods Section, American Statistical Association, Washington, DC, 3257–3264.

LOHR, S.L. (2013) *Dual frame surveys: Recent developments and challenges*, http://new.sis-statistica.org/wp-content/uploads/2013/09/RS10-A-Generalized-Composite-Index-based-on-Non-substitutability-of-Individual-Indicators.pdf

LOHR, S.L. & RAO, J.N.K. (2000) *Inference from dual frame surveys*, J. Amer. Statist. Assoc. **94**, 271–280.

LOHR, S.L. & RAO, J.N.K. (2006) *Estimation in multiple frame surveys*, J. Amer. Statist. Assoc. **101**, 1019–1030.

LUMLEY, T. & SCOTT, A.J. (2013) *Partial likelihood ratio tests for the Cox model under complex sampling*, Stat. Med. **31**, 110–123.

LUMLEY, T. & SCOTT, A.J. (2014) *Tests for regression models fitted to survey data*, Aust. N. Z. J. Stat. **56**(1), 1–14.

METCALF, P. & SCOTT, A.J. (2009) *Using multiple frames in health surveys*, Statist. Med. **28**, 1512–1523.

SKINNER, C.J. (1991) *On the efficiency of raking ratio estimation for multiple frame surveys*, J. Amer. Statist. Assoc. **86**, 779–784.

SKINNER CJ, RAO J.N.K. (1996) *Estimation in dual frame surveys with complex designs*, J. Amer. Statist. Assoc. **91**, 349–356.

RAO J.N.K. (2011) *Impact of Frequentist and Bayesian Methods on Survey Sampling Practice: A Selective Appraisal*, Statistical Science **26**(2), 240–256.