On the Allocation and Estimation for Dual Frame Survey Data

A. Demnati[1], J. N. K. Rao[2], M. A. Hidiroglou[3], and J.-L. Tambay[4]

A. Demnati, Social Survey Methods Division, Statistics Canada, Ottawa, Canada[1]
J. N. K. Rao, School of Mathematics and Statistics, Carleton University, Ottawa, Canada[2]
M. A. Hidiroglou, Research and Innovation Methodology Division, Statistics Canada, Ottawa, Canada[3]
J.-L. Tambay, Household Survey Methods Division, Statistics Canada, Ottawa, Canada[4]

Abstract

With increase in the number of surveys, the cost of personal interviewing using a complete frame has increased significantly. As a result, new surveys are often conducted using dual frames with one frame or both frames cheaper to sample but incomplete. Under simple random sampling in both frames, we consider the determination of "optimal" frame sample sizes that minimize the cost subject to constraints on the variances of dual frame estimators of totals for one or more characteristics of interest. The case of estimators calibrated to known frame sizes is also studied. Dual frame estimators based on multiple weight adjustments to account for nonresponse, multiplicities, and calibration to know auxiliary totals are also given. Finally, we apply the Demnati and Rao (2004) method to take account of such multiple weight adjustments in variance estimation.

Keywords: Calibration, incomplete frame, multiplicity, optimal sample size, variance estimation.

1. Introduction

In a dual frame survey, samples are drawn independently from two frames F_1 and F_2. We assume that frames F_1 and F_2 together cover the population of interest, F. In one example, one frame is complete, say $F_1 = F$, but is expensive to sample, whereas the other frame F_2 is incomplete but cheap to sample. Hartley (1962, 1974) demonstrated the advantages of sampling both frames in this case to arrive at more efficient estimators for the same cost compared to sampling from the complete frame only. In another example, both frames F_1 and F_2 are incomplete: F_1 is a frame of landline telephones and F_2 is a frame of cellular telephone numbers (Lohr and Rao, 2006). We were motivated by the following application. Consider a fixed period of traffic during which a collection of trips moves through a network. Each trip originates at one node in the network and travels to another node along a path. A survey is conducted to produce a profile of the volume and characteristics of the network by taking random samples on each directed link or site. For example, the 1999 National Roadside Study conducted roadside observations and interviews on more than 250 sites (directed links) to produce a profile of the volume and characteristics of the trucking activity in Canada. The road network covers more than 25 thousand kilometers of mostly the National Highway System, augmented by routes of regional importance to trucking. The survey period is one week in order to capture day and hour variation. The data collected at each site consists, in part, of a random sample of interviews and observations of the trips, and in another part, of a series of count of trucks passing the site during the survey period. Data collected from different sites are integrated into a single data set. This integration can be easily expressed in terms of multiple frame where each site represents an incomplete frame of trips population. By identifying the route of each trip, we can determine the multiplicity of each trip, i.e., the number of sites (frames) reporting a given trip.

In this paper, we study two problems in dual frame surveys. In section 2, we consider simple random sampling in both frames and obtain "optimal" frame sample sizes, n_1 and n_2, that minimize the cost subject to constraints on the variances of dual frame estimators of totals for one or more characteristics of interest. We obtain optimal n_1 and n_2 for the dual frame estimator of Hartley (1962) as well as the "single" frame estimators proposed by Kalton and Anderson (1986) (also Skinner, 1991) and Bankier (1986). The case of calibration to known frame sizes N_1 and N_2 is also studied. Section 3 shows how to account for nonresponse, multiplicities and calibration to know auxiliary totals through multiple weight adjustments. Finally, variance estimation under multiple weight adjustments is studied in section 4, using the Demnati and Rao (2004) linearization method.

2. Determination of optimal sample sizes

2.1 Hartley's dual frame estimator

In dual frame surveys, the population total Y of a characteristic of interest y can be expressed as

$$Y = \sum_f \sum_k J_{fk} y_k \phi_{fk}, \qquad (2.1)$$

where $\sum_f$ represents summation over the frames, $f = 1,2$, $\sum_k$ represents summation over population elements, J_{fk} is the frame f membership indicator variable for element k, $\phi_{1k} = \varphi$ and $\phi_{2k} = 1-\varphi$ if element k is in both frame with $0 \le \varphi \le 1$, and $\phi_{fk} = 1$ if element k is only in one frame. We assume that samples are independently drawn from each frame. The basic design weights for frame f are given by

$$d_{fk} = J_{fk} a_{fk} / \pi_{fk}, \qquad (2.2)$$

where a_{fk} is the conditional sample membership indicator for element k in frame f and $\pi_{fk} = E(a_{fk} \mid J_{fk} = 1)$ is the conditional probability of selection of element k from frame f. Hartley's (1962) dual frame unbiased estimator of the total Y is given by

$$\hat{Y}_H = \sum_f \sum_k d_{fk} \phi_{fk} y_k = \sum_f \hat{Y}_f^*, \qquad (2.3)$$

where $\hat{Y}_f^* = \sum_k d_{fk} y_{fk}^*$ and $y_{fk}^* = \phi_{fk} y_k$.

The sampling variance of $\hat{Y}_H$ is

$$Var(\hat{Y}_H) = \sum_f Var(\hat{Y}_f^*) \equiv \sum_f V_f(y^*), \qquad (2.4)$$

where $V_f(u)$, in operation notation, is the sampling variance of the estimated total $\hat{U}_f = \sum_k u_k d_{fk}$ for frame f.

Under simple random sampling (SRS) in both frames we have

$$V_f(u) = N_f^2 (1 - n_f / N_f) S_f^2(u) / n_f, \qquad (2.5)$$

where $S_f^2(u) = \sum_k J_{fk} (u_k - \overline{U}_f)^2 / (N_f - 1)$ with $\overline{U}_f = \sum_k J_{fk} u_k / N_f$ and n_f is the sample size from frame f.

Suppose we consider p characteristics of interest $y_1,..., y_p$. Then, under SRS, it follows from (2.4) and (2.5) that for a specified φ we can express $Var(\hat{Y}_{Hj})$ for the j^{th} variable y_j as

$$Var(\hat{Y}_{Hj}) = v_{j0} + \sum_f v_{jf} / n_f, \; j = 1,..., p \qquad (2.6)$$

where $v_{j0} = -\sum_f N_f S_f^2(y_j^*)$, $v_{jf} = N_f^2 S_f^2(y_j^*)$, and $y_{jfk}^* = y_{jk} \phi_{fk}$. We first determine the optimal n_1 and n_2 for a specified φ such that the cost

$$C = c_0 + \sum c_f n_f \qquad (2.7)$$

is minimized subject to constraints on the p variances:

$$Var(\hat{Y}_{Hj}) \le V_j, \; j = 1,..., p \qquad (2.8)$$

where c_0 is the fixed cost, c_f is the cost per unit in frame f and the V_j are specified tolerances. For example, one could specify upper limits, δ_j, on the coefficient of variation of $\hat{Y}_{Hj}$ so that $V_j = (\delta_j Y_j)^2$. We can improve the efficiency of the unbiased estimator $\hat{Y}_H$ by calibrating on the known sizes N_1 and N_2 (Bankier, 1986). In particular, a generalized regression estimator (GREG), $\tilde{Y}_H$, can be used to ensure calibration to N_1 and N_2. In case of GREG, we replace (2.8) by

$$Var(\tilde{Y}_{Hj}) \le V_j, \; j = 1,..., p. \qquad (2.9)$$

It is easily seen that

$$Var(\tilde{Y}_H) \approx \sum_f V_f(e^*), \qquad (2.10)$$

where

$$e_{fk}^* = \phi_{fk}(y_k - \boldsymbol{t}_k^T \boldsymbol{B}), \qquad (2.11)$$

with $\boldsymbol{T} = (N_1, N_2)^T$, $\boldsymbol{t}_k = (J_{1k}, J_{2k})^T$ and $\boldsymbol{B} = [\sum_f \sum_k \phi_{fk} \boldsymbol{t}_k \boldsymbol{t}_k^T]^{-1} \sum_f \sum_k \phi_{fk} \boldsymbol{t}_k y_k$. By letting $x_f = n_f^{-1}$, the cost C becomes a separable convex function in the x_f and the constants (2.8) or (2.9) change to linear functions of the x_f. Hence, the optimization problem is reduced to a standard convex programming problem. The optimal φ and associated n_1 and n_2 can be obtained by minimizing the optimal cost $C(\varphi)$ with respect to φ.

Note that the unbiased estimator $\hat{Y}_H$ given by (2.3) with the optimal φ uses a common weight for all variables y and ensures that the constraints (2.8) are satisfied for the variables $y_1,..., y_p$ with minimum cost. Thus is also true for the GREG $\tilde{Y}_H$.

Example

We generated a population $\{(y_{1k}, y_{2k}, y_{3k}, y_{4k})\}$ of size $N = 1{,}000$, where $y_{1k} \sim B(1, 0.6)$, $y_{2k} = 50 + 16 \times \varepsilon_k$ with $\varepsilon_k \sim N(0,1)$, $y_{3k} \sim B(1, p_k)$, with $p_k = \exp(0.1 + 1 \times J_{2k}) / (1 + \exp(0.1 + 1 \times J_{2k}))$, and $y_{4k} = 50 + J_{2k} \times 50 + 4 \times \varepsilon_{1k} + J_{2k} \times 10 \times \varepsilon_{2k}$ with

$\varepsilon_{1k} \sim N(0,1)$ and $\varepsilon_{2k} \sim N(0,1)$.

The above choice of p_k, gives $p_k \approx 0.75$ when $J_{2k} = 1$ and $p_k \approx 0.52$ when $J_{2k} = 0$.

Frame 1 membership indicator is set to $J_{1k} = 1$, which mean that frame 1 is complete (as the case of an area frame), and frame 2 membership indicator variable is generated from $J_{2k} \sim B(1,0.6)$, which assume a 60% coverage of frame 2. We assume simple random sampling to be used in each frame. For the cost, we set $c_0 = 0$, $c_1 = 1$ and two different costs are used for c_2: 0.5 and 0.2. We set $\delta_j = 0.05$, $j = 1,...,4$, for the tolerances. Table1 reports the multivariate optimization results for $\varphi = 0.5$ using both the basic estimator $\hat{Y}_H$ and the GREG estimator $\tilde{Y}_H$. We have also included the results in the case of sampling only from the complete frame 1. First, Table 1 shows that we reduce the minimum cost for a given c_2 by using GREG: with $c_2 = .2$, $C_{\min} = 165$ for $\tilde{Y}_H$ compared to 188 for $\hat{Y}_H$. Secondly, we note that the minimum cost, $C_{\min}$, for the dual frame approach goes down as c_2 decreases: for the GREG $C_{\min} = 196$ with $c_2 = .5$ compared to 165 with $c_2 = .2$. Third, it is interesting to note that $C_{\min}$ for dual frames can be larger than the $C_{\min}$ using only the complete frame if c_2 / c_1 is not small: with $c_2 = .5$, $C_{\min} = 220$ for $\hat{Y}_H$ compared to $C_{\min} = 203$ for the single complete frame estimator. However, as c_2 / c_1 decreases, use of dual frames can lead to significant reduction in the minimum cost using the GREG: $C_{\min} = 165$ compared to $C_{\min} = 203$ for the complete frame only estimator.

To determine the optimal value for φ, we repeated the optimization process for different value of φ between 0 and 1, and the results are given in Figure 1. The resulting optimal value for φ, n_1, n_2 and $C_{\min}$ are reported in Table 2. Comparing the results in Tables 1 and 2, we see that $C_{\min}$ is somewhat reduced by using the optimal φ relative to $\varphi = 0.5$: for the GREG with $c_2 = .2$, $C_{\min} = 158$ using $\varphi_{opt} = 0.22$ compared to $C_{\min} = 165$ using $\varphi = 0.5$. However, $C_{\min}(\varphi)$ seems to be fairly flat near φ_{opt} (see Figure 1).

2.2 "Single" frame estimators

In some cases, dual frame surveys are treated as single frame surveys by combining the two samples. Kalton and Anderson (1986) and Skinner (1991) proposed a "single" frame estimator,

$$\hat{Y}_S = \sum_f \sum_k d_{fk}(1 - I_k) y_k + \sum_f \sum_k d_{fk} I_k y_k \phi_{fk} \quad (2.12)$$

for general designs in the two frames, where I_k is the overlap membership indicator for element I_k and

$$\phi_{fk} = \frac{\pi_{fk}}{(\pi_{1k} + \pi_{2k})}. \quad (2.13)$$

We can improve the efficiency of $\hat{Y}_S$ by calibrating on the known frame sizes. Denote the resulting GREG as $\tilde{Y}_S$. Let $x_k = (1 - I_k) y_k$, $z_k = I_k y_k$ then the variance of $\hat{Y}_S$ is given by

$$\begin{aligned} Var(\hat{Y}_S) = &\sum_f Var(\sum_k d_{fk} x_k) \\ &+ \sum_f Var(\sum_k d_{fk} z_k \phi_{fk}) \\ &+ 2\sum_f Cov(\sum_k d_{fk} x_k, \sum_k d_{fk} z_k \phi_{fk}). \end{aligned} \quad (2.14)$$

Approximate variance of GREG $\tilde{Y}_S$ is obtained by changing y_k to $y_k - \boldsymbol{t}_k^T \boldsymbol{B}$ in (2.14).

Under SRS in each frame, we have

$$\phi_{1k} = \phi_1 = \frac{n_1 N_2}{n_1 N_2 + n_2 N_1}, \ \phi_2 = 1 - \phi_1,$$

$$\sum_f Var(\sum_k d_{fk} x_k) = \sum_f N_f (N_f / n_f - 1) S_{f;xx},$$

$$\sum_f Var(\sum_k d_{fk} z_k \phi_{fk}) = \sum_f \phi_f^2 N_f (N_f / n_f - 1) S_{f;zz}$$

and

$$\sum_f Cov(\sum_k d_{fk} x_k, \sum_k d_{fk} z_k \phi_{fk}) = \sum_f \phi_f N_f (N_f / n_f - 1) S_{f;xz}$$

where $S_{f;xz} = \sum J_{fk} (x_k - \bar{X}_f)(z_k - \bar{Z}_f)/(N_f - 1)$ and $\bar{Z}_f = \sum J_{fk} z_k / N_f$.

The allocation problem consists of minimizing the cost of the survey given by (2.7), subject to sampling variance constraints

$$Var(\hat{Y}_{Sj}) \le V_j, \ j = 1,..., p \quad (2.15)$$

or

$$Var(\tilde{Y}_{Sj}) \le V_j, \ j = 1,..., p \quad (2.16)$$

with $Var(\hat{Y}_{Sj})$ and $Var(\tilde{Y}_{Sj})$ for $j = 1,..., p$ obtained from (2.14). The variances $Var(\hat{Y}_{Sj})$ and $Var(\tilde{Y}_{Sj})$ do not have the separable form (2.6), but non-linear programming can be used to determine the optimal n_1 and n_2.

Bankier (1986) removed the duplicate sampled units in the overlap domain and proposed a Horvitz-Thompson (HT) estimator

$$\hat{Y}_B = \sum_k d_k y_k , \quad (2.17)$$

as the unbiased estimator of the population total Y, where $\sum_k$ is the sum over all the distinct units in the combined sample, $d_k = a_k / \pi_k$, $a_k = 1 - \prod_f (1 - a_{fk})$, $\pi_k = E(a_k) = 1 - \prod_f (1 - \pi_{fk})$. Denoting the corresponding GREG that calibrates to N_1 and N_2 as $\widetilde{Y}_B$.

The variance of $\hat{Y}_B$ is given by the well known HT variance formula

$$Var(\hat{Y}_B) = \sum y_k^2 (\pi_k^{-1} - 1) + 2 \sum_k \sum_{l<k} y_k y_l (\pi_{kl} / (\pi_k \pi_l) - 1), \quad (2.18)$$

with for $k \neq l$

$$\pi_{kl} = \pi_k + \pi_l - 1 + \prod_f (1 - \pi_{fk})(1 - \pi^*_{fl}), \quad (2.19)$$

and $\pi^*_{fl} = \Pr(l \in s_f \mid k \notin s_f)$. Appropriate variance of $\widetilde{Y}_B$ is obtained by changing y_k to $y_k - \boldsymbol{t}_k^T \boldsymbol{B}$ in (2.18). Again $Var(\hat{Y}_{Bj})$ and $Var(\widetilde{Y}_{Bj})$ do not have the separable form (2.6), but non-linear programming can be used to determine the optimal n_1 and n_2 that minimize the cost (2.7) subject to $Var(\hat{Y}_{Bj}) \le V_j$ or $Var(\widetilde{Y}_{Bj}) \le V_j$, $j = 1,..., p$.

Example (continuation)

For the example in section 2.1, Table 3 and 4 report the optimal n_1, n_2 and minimum cost ($C_{\min}$) for the Kalton-Anderson estimator and the Bankier estimator, respectively. From Tables 3 and 4, we note that GREG leads to significant reduction in minimum cost when c_2 / c_1 is small ($c_2 = 0.2$): $C_{\min} = 159$ for $\widetilde{Y}_S$ compared to 190 for $\hat{Y}_S$; $C_{\min} = 155$ for $\widetilde{Y}_B$ compared to 194 for $\hat{Y}_B$. Using GREG, the Bankier estimator leads to slightly lower cost compared to the Kalton-Anderson estimator: for $c_2 = 0.2$, $C_{\min} = 155$ for $\widetilde{Y}_B$ compared to $C_{\min} = 159$ for $\widetilde{Y}_S$. It is interesting to note that $C_{\min} = 155$ for $\widetilde{Y}_B$ is slightly smaller than $C_{\min} = 158$ for $\widetilde{Y}_H$ with optimal φ (Table 2) because the duplicate sampled units in the overlap domain are not removed in the case of $\widetilde{Y}_H$.

3. Multiple Weight Adjustments

In the presence of missing responses, weighting adjustment is often used to compensate for unit (or complete) nonresponse. Let r_{fk} denotes the partial response indicator variable for element k in frame f, i.e. $r_{fk} = 0$ if there is complete nonresponse and $r_{fk} = 1$ if there is partial response.

A widely-used approach to adjust for unit nonresponse in each frame, when predictor variables $\boldsymbol{x}_{fk} = (x_{1fk},..., x_{q_f fk})^T$ are available for all sampled elements, is to use the GREG calibration weights (Lundström and Särndal, 1999):

$$w_{fk}^{(1)} = \breve{d}_{fk} g_{fk}^{(1)},$$

where

$$\breve{d}_{fk} = d_{fk} r_{fk}, \quad (3.1)$$

the “g-weights” are given by

$$g_{fk}^{(1)} = 1 + (\hat{\boldsymbol{X}}_f - \breve{\boldsymbol{X}}_{fr})^T [\sum_k \breve{d}_{fk} c_{fk}^{(1)} \boldsymbol{x}_{fk} \boldsymbol{x}_{fk}^T]^{-1} c_{fk}^{(1)} \boldsymbol{x}_{fk}, \quad (3.2)$$

for specified $c_{fk}^{(1)}$, $\hat{\boldsymbol{X}}_f = \sum_k d_{fk} \boldsymbol{x}_{fk}$ is the HT estimator of the frame f total X_f of the $q_f \times 1$ vector $\boldsymbol{x}_{fk}$, and $\breve{\boldsymbol{X}}_{fr} = \sum_k \breve{d}_{fk} \boldsymbol{x}_{fk}$ is the HT estimator of the frame f respondent total $\boldsymbol{X}_{fr}$ of the vector $\boldsymbol{x}_{fk}$. The resulting GREG estimator of the frame f total Y_f, namely $\hat{Y}_f = \sum_k w_{fk}^{(1)} y_k$, has the calibration property

$$\sum_k w_{fk}^{(1)} \boldsymbol{x}_{fk} = \hat{\boldsymbol{X}}_f . \quad (3.3)$$

Note that the right side of (3.3) is a random variable.

A common approach to handle unit nonresponse is to classify respondents and non respondents into q_f adjustment classes, using auxiliary information on all sample elements, in which case x_{cfk} denotes the group c, $c = 1,..., q_f$, membership indicator variable for element k with $\sum_c x_{cfk} = 1$. In this case, the GREG adjustment factor given by (3.2) with $c_{fk}^{(1)} = 1$ reduces to

$$g_{fk}^{(1)} = (\hat{N}_f^{(1)} / \breve{N}_{fr}^{(1)},..., \hat{N}_f^{(q_f)} / \breve{N}_{fr}^{(q_f)}) \boldsymbol{x}_{fk},$$

where $(\hat{N}_f^{(1)},..., \hat{N}_f^{(q_f)})$ is the vector estimate of the class sizes and $(\breve{N}_{fr}^{(1)},..., \breve{N}_{fr}^{(q_f)})$ is the vector estimates of the respondent class sizes.

When aggregating the samples from the two frames, a second adjustment has to be made to account for the multiplicity of each element:

$$w_{fk}^{(2)} = w_{fk}^{(1)} \phi_{fk} .$$

Suppose an additional vector of calibration variables $\boldsymbol{t}_k = (t_{1k},...,t_{qk})^T$ with know totals $\boldsymbol{T} = (T_1,...,T_q)^T$ is available in addition to the vectors $\boldsymbol{x}_{fk}$. The vectors $\boldsymbol{x}_{fk}$ are assumed to be related to the response probability of element k, while the vector $\boldsymbol{t}_k$ is assumed to be related to the variables of interest. In this case, the final GREG calibration weights w_{fk} are given by

$$w_{fk}^{(3)} = w_{fk}^{(2)} g_{fk}^{(2)},$$

where the "g-weights" are given by

$$g_{fk}^{(2)} = 1 + (\boldsymbol{T} - \hat{\boldsymbol{T}}^{(2)})^T [\textstyle\sum_f \sum_k w_{fk}^{(2)} c_{fk}^{(2)} \boldsymbol{t}_{fk} \boldsymbol{t}_{fk}^T]^{-1} c_{fk}^{(2)} \boldsymbol{t}_{fk}, \quad (3.4)$$

for specified $c_{fk}^{(2)}$, and $\hat{\boldsymbol{T}}^{(2)} = \sum_f \sum_k w_{fk}^{(2)} \boldsymbol{t}_{fk}$. The resulting GREG estimator of the population total Y, namely $\hat{Y} = \sum_f \sum_k w_{fk}^{(3)} y_k$ has the calibration property

$$\textstyle\sum_f \sum_k w_{fk}^{(3)} \boldsymbol{t}_{fk} = \boldsymbol{T}. \quad (3.5)$$

4. Demnati-Rao Linearization Method

After adjustment for complete nonresponse, multiplicities, and use of auxiliary information, the estimator $\hat{Y}$ of Y is given by

$$\hat{Y} = \textstyle\sum_f \sum_k \breve{d}_{fk} g_{fk}^{(1)} \phi_{fk} g_{fk}^{(2)} y_k, \quad (4.1)$$

where $\breve{d}_{fk}$ is defined by (3.1), $g_{fk}^{(1)}$ is defined by (3.2) and $g_{fk}^{(2)}$ is define by (3.4). Let $\boldsymbol{d}_k = (\boldsymbol{d}_{1k}^T, \boldsymbol{d}_{2k}^T)^T$, $\boldsymbol{d}_{fk} = (d_{1fk}, d_{2fk})^T$ with $d_{1fk} = d_{fk}$ and $d_{2fk} = d_{fk} r_{fk}$. It follows from (4.1) that $\hat{Y}$ is of the form $\boldsymbol{f}(\boldsymbol{A}_d)$ where $\boldsymbol{A}_d$ is a $4 \times N$ matrix with k^{th} column $\boldsymbol{d}_k$. In operator notation, let $\vartheta(\boldsymbol{u})$ denote the estimator of total variance of a linear estimator $\hat{U} = \sum\sum \boldsymbol{u}_{fk}^T \boldsymbol{d}_{fk}$. Then, Demnati and Rao (2007) have shown that a linearization variance estimator of $\hat{Y} = f(\boldsymbol{A}_d)$ is simply given by

$$\vartheta_{DR}(\hat{Y}) = \vartheta(\boldsymbol{z}), \quad (4.2)$$

where $\vartheta(\boldsymbol{z})$ is obtained from $\vartheta(\boldsymbol{u})$ by replacing $\boldsymbol{u}_k$ by $\boldsymbol{z}_k = \partial f(\boldsymbol{A}_b)/\partial \boldsymbol{b}_k |_{\boldsymbol{A}_b = \boldsymbol{A}_d}$, where $\boldsymbol{A}_b$ is a $4 \times N$ matrix of arbitrary real numbers with k^{th} column $\boldsymbol{b}_k$. Following the explicit differentiation method of Demnati and Rao (2004), $\boldsymbol{z}_k = \partial f(\boldsymbol{A}_b)/\partial \boldsymbol{b}_k |_{\boldsymbol{A}_b = \boldsymbol{A}_d} \equiv (\boldsymbol{z}_{1k}^T, \boldsymbol{z}_{2k}^T)$ is evaluated as:

$$\boldsymbol{z}_{fk} = (\boldsymbol{z}_{1fk}, \boldsymbol{z}_{2fk})^T, \quad (4.3)$$

with $z_{1fk} = \hat{\boldsymbol{B}}_f^T(e_f^*) \boldsymbol{x}_{fk}$,

$$z_{2fk} = g_{fk}^{(1)} \left(e_{fk}^* - \hat{\boldsymbol{B}}_f^T(e_f^*) \boldsymbol{x}_{fk} \right),$$

where

$$e_{fk}^* = \phi_{fk} g_{fk}^{(2)} \left(y_k - \hat{\boldsymbol{B}}^T(y) \boldsymbol{t}_k \right),$$

$$\hat{\boldsymbol{B}}_f(e_f^*) = [\textstyle\sum_k d_{fk} r_{fk} c_{fk}^{(1)} \boldsymbol{x}_{fk} \boldsymbol{x}_{fk}^T]^{-1} \sum_k d_{fk} r_{fk} c_{fk}^{(1)} \boldsymbol{x}_{fk} e_{fk}^*,$$

and

$$\hat{\boldsymbol{B}}(y) = [\textstyle\sum_f \sum_k w_{fk}^{(2)} c_{fk}^{(2)} \boldsymbol{t}_{fk} \boldsymbol{t}_{fk}^T]^{-1} \sum_f \sum_k w_{fk}^{(2)} c_{fk}^{(2)} \boldsymbol{t}_{fk} y_k.$$

It remains to evaluate $\vartheta(\boldsymbol{u})$. We have

$$\vartheta(\boldsymbol{u}) = \textstyle\sum_{fk} \sum_{gt} \boldsymbol{u}_{fk}^T \operatorname{cov}(\boldsymbol{d}_{fk}, \boldsymbol{d}_{gt}) \boldsymbol{u}_{gt}, \quad (4.4)$$

with

$$\begin{aligned} \operatorname{cov}(\boldsymbol{d}_{fk}, \boldsymbol{d}_{gt}) = {} & d_{kt}^{(fg)} r_{fk} r_{gt} [(\hat{\xi}_{kt}^{(fg)} - \hat{\xi}_{fk} \hat{\xi}_{gt}) / \hat{\xi}_{kt}^{(fg)}] \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \\ & + d_{kt}^{(fg)} [(1 - \omega_{kt}^{(fg)}) / \omega_{kt}^{(fg)}] \boldsymbol{v}_{fk} \boldsymbol{v}_{gt}^T, \end{aligned} \quad (4.5)$$

where $\sum_{fk} = \sum_f \sum_k$, $\boldsymbol{v}_{fk} = (1, r_{fk})^T$, $\hat{\xi}_{fk} = \hat{E}_r(r_{fk})$, $\hat{\xi}_{kt}^{(fg)} = \hat{E}_r(r_{fk} r_{gt})$, $d_{kt}^{(fg)} = d_{fk} d_{gt} / E[d_{fk} d_{gt}]$, $\omega_{kt}^{(fg)} = E[d_{fk}] E[d_{gt}] / E[d_{fk} d_{gt}]$ and E_r is the response expectation. If $f \neq g$ then $E[d_{fk} d_{gt}] = 1$, $d_{kt}^{(fg)} = d_{fk} d_{gt}$, and $\omega_{kt}^{(fg)} = 1$.

Substituting $\boldsymbol{z}_k$ in (4.3) for $\boldsymbol{u}_k$ in (4.4), we get

$$\begin{aligned} \vartheta_{DR}(\hat{Y}) = {} & \textstyle\sum_{fk} \sum_{gt} d_{kt}^{(fg)} r_{fk} r_{gt} [(\hat{\xi}_{kt}^{(fg)} - \hat{\xi}_{fk} \hat{\xi}_{gt}) / \hat{\xi}_{kt}^{(fg)}] z_{fk;r} z_{gt;r} \\ & + \textstyle\sum_{fk} \sum_{gt} d_{kt}^{(fg)} [(1 - \omega_{kt}^{(fg)}) / \omega_{kt}^{(fg)}] z_{fk;s} z_{gt;s} \\ \equiv {} & \vartheta_r + \vartheta_s \end{aligned} \quad (4.6)$$

where $z_{fk;r} = g_{fk}^{(1)} \left(e_{fk}^* - \boldsymbol{B}_f^T(e_f^*) \boldsymbol{x}_{fk} \right)$, and $z_{fk;s} = r_{fk} g_{fk}^{(1)} \left(e_{fk}^* - \hat{\boldsymbol{B}}_f^T(e_f^*) \boldsymbol{x}_{fk} \right) + \hat{\boldsymbol{B}}_f^T(e_f^*) \boldsymbol{x}_{fk}$.

Note that the first component, ϑ_r, corresponds to the response mechanism and the second component, ϑ_s, corresponds to the sampling design.

Under simple random sampling in each frame,

$$\vartheta_s = \textstyle\sum_f N_f^2 (1 - n_f / N_f) / n_f s_{fz_s}^2, \quad (4.7)$$

where $s_{fx}^2 = \sum a_{fk} (x_{fk} - \bar{x}_f)^2 / (n_f - 1)$ and $\bar{x}_f = \sum a_{fk} x_{fk} / n_f$.

Under independent response mechanism

$$\begin{aligned} \vartheta_r = {} & \textstyle\sum_f \sum_k d_{fk} r_{fk} (1 - \hat{\xi}_{fk}) z_{fk;r} z_{fk;r} \\ & + 2 \textstyle\sum_k \sum_t d_{1k} d_{2t} r_{1k} r_{2t} 1(k = t) [(\hat{\xi}_{kt}^{(12)} - \hat{\xi}_{1k} \hat{\xi}_{2t}) / \hat{\xi}_{kk}^{(12)}] z_{1k;r} z_{2t;r}, \end{aligned} \quad (4.8)$$

where $1(k = t) = 1$ if element k is the same as element t and $1(k = t) = 0$ if not.

The sum of (4.7) and (4.8) constitutes $\vartheta_{DR}(\hat{Y})$.

We conducted a small simulation study to examine the unconditional (design-response) performances of ratio estimator $\hat{Y}_R^{(3)}$ of the finite population total $\theta_N = Y$. In particular, we compared the efficiency of $\hat{Y}_R^{(3)}$, using the three weight adjustments, relative to $\hat{Y}_R^{(2)}$ using only the adjustments for nonresponse and multiplicities. Here $\hat{Y}_R^{(.)} = X\hat{Y}^{(.)} / \hat{X}^{(.)}$ where $\hat{Y}^{(.)} = \sum\sum w_{fk}^{(.)} y_k$. We also examined the unconditional performance of the variance estimators $\vartheta_{DR}(\hat{Y}_R^{(3)})$ and $\vartheta_{DR}(\hat{Y}_R^{(2)})$ in tracking the total variances of $\hat{Y}_R^{(3)}$ and $\hat{Y}_R^{(2)}$, respectively. Note that $\vartheta_{DR}(\hat{Y}_R^{(.)})$ is given by $\vartheta_{DR}(\sum\sum u_{fk} w_{fk}^{(.)})$ where $u_{fk} = X(y_k - \hat{Y}^{(.)} / \hat{X}^{(.)}) / \hat{X}^{(.)}$. We first generated one finite populations $\{y_k\}$ of size $N = 393$, from the ratio model

$$y_k = 2\, x_k + x_k^{1/2} \varepsilon_k ,$$

with ε_k are independent observations generated from a $N(0,1)$, where the fixed x_k are the "number of beds" for the Hospitals population studied in Valliant *et al*. (2000, p.424-427). We set $J_{1k} = 1$ and we generate J_{2k} from $B(p_k)$ where $logit(p_k) = 1 - 0.003 x_k$. This choice gives $N_1 = 393$ and $N_2 = 189$. We set $(c_0, c_1, c_2) = (0, 1, 0.5)$ and $\delta = .05$. Using Kalton and Anderson estimator, the optimal simple random sample sizes are $n_1 = 104$ and $n_2 = 55$. In order to set up the response mechanism, we first grouped population units into two classes: class 1 constitutes units k having $x < 350$, and class 2 constitutes units having $x \geq 350$. The response probabilities are set as follows: Frame 1: 0.70 for class 1, and 0.90 for class 2. Frame 2: 0.60 for class 1, and 0.80 for class 2. From the two frames, we generated $R = 20{,}000$ dual frames with nonresponses. From each generated dual frame, one SRS of size 104 was drawn from frame 1, and one SRS of size 55 was drawn from frame 2. We calculated the ratio estimates $\hat{Y}_R^{(2)}$, $\hat{Y}_R^{(3)}$, and the variance estimates $\vartheta_{DR}(\hat{Y}_R^{(2)})$, $\vartheta_{DR}(\hat{Y}_R^{(3)})$, from each combined sample and their means $\overline{\hat{Y}}_R^{(2)}$, $\overline{\hat{Y}}_R^{(3)}$, $\overline{\vartheta}_{DR}(\hat{Y}_R^{(2)})$, and $\overline{\vartheta}_{DR}(\hat{Y}_R^{(3)})$, and the variance of $\hat{Y}_R^{(2)}$ and $\hat{Y}_R^{(3)}$, denoted $V(\hat{Y}_R^{(2)})$ and $V(\hat{Y}_R^{(3)})$. We have the following results:

(1) $V(\hat{Y}_R^{(3)}) / V(\hat{Y}_R^{(2)}) = 1.0113$, suggesting that post-stratification is not effective with ratio estimation when the model fits the data well; in fact, it lead to slight increase in variance. This result is in agreement with the observation made by Rao, Yung, and Hidiroglou (2002).

(2) The relative biases of DR variance estimators are: $(\overline{\vartheta}_{DR}(\hat{Y}_R^{(2)}) - V(\hat{Y}_R^{(2)})) / V(\hat{Y}_R^{(2)}) = -2.9\%$ and $(\overline{\vartheta}_{DR}(\hat{Y}_R^{(3)}) - V(\hat{Y}_R^{(3)})) / V(\hat{Y}_R^{(3)}) = -3.8\%$, showing that ϑ_{DR} tracks the total variance with two or three weight adjustments.

References

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

Demnati, A. and Rao, J.N.K. (2004), "Linearization Variance Estimators for Survey Data (with discussion)," *Survey Methodology*, 30, 17-34.

Demnati, A. and Rao, J.N.K. (2007), "Linearization Variance Estimators for Survey Data: Some Recent Work," *Third International Conference on Establishment Surveys*, Montréal, Canada.

Hartley, H. O. (1962), "Multiple Frame Surveys," *in Proceedings of the Social Statistics Section, American Statistical Association*, 203-206.

Hartley, H. O. (1974),"Multiple Frame Methodology and Selected Application," *Sankhyā*, Ser. C, Part 3, 36, 99-118.

Kalton, G., and Anderson, D.W. (1986),"Sampling Rare Populations," *Journal of the Royal Statistical Society,* Ser. A, 149, 65-82.

Lohr, S., and Rao, J.N.K. (2006), "Estimation in Multiple-Frame Surveys," *Journal of the American Statistical Association,* 101, 1019-1030.

Lundström, S., and Särndal, C.-E., (1999), "Calibration as a Standard Method for Treatment of Nonresponse," *Journal of Official Statistics*, 15, 305-327.

Rao, J.N.K., Yung, W. and Hidiroglou, M. (2002), "Estimating Equations for the Analysis of Survey Data using Poststratification Information", *Sankyhā: The Indian Journal of Statistics*, 64, 1-15.

Skinner, C.J. (1991), "On the Efficiency of Raking Ratio Estimation for Multiple Frame Surveys," *Journal of the American Statistical Association,* 86, 779-784.

Valliant R., Dorfman, A. H. and Royall, R. M. (2000) "*Finite Population Sampling and Inference: A Prediction Approach*", Wiley.

Table 1: Optimal n_1, n_2 and C_{min} using $\varphi = 0.5$: Hartley's estimator

	Estimator	n_1	n_2	$C_{\min}$
Complete frame		203		203
$c_2 = 0.5$	Basic	174	92	220
	GREG	152	88	196
$c_2 = 0.2$	Basic	161	134	188
	GREG	140	127	165

Table 2: Optimal φ, n_1, n_2 and C_{min} : Hartley's estimator

	Estimator	φ	n_1	n_2	$C_{\min}$
$c_2 = 0.5$	Basic	.87	190	23	202
	GREG	.57	158	75	196
$c_2 = 0.2$	Basic	.64	166	96	186
	GREG	.22	119	191	158

Table 3: Optimal n_1, n_2 and C_{min} : Kalton-Anderson estimator

	Estimator	n_1	n_2	$C_{\min}$
$c_2 = 0.5$	Basic	197	12	203
	GREG	157	79	197
$c_2 = 0.2$	Basic	173	83	190
	GREG	121	187	159

Table 4: Optimal n_1, n_2 and C_{min} : Bankier estimator

	Estimator	n_1	n_2	$C_{\min}$
$c_2 = 0.5$	Basic	201	4	203
	GREG	151	78	190
$c_2 = 0.2$	Basic	181	64	194
	GREG	119	176	155

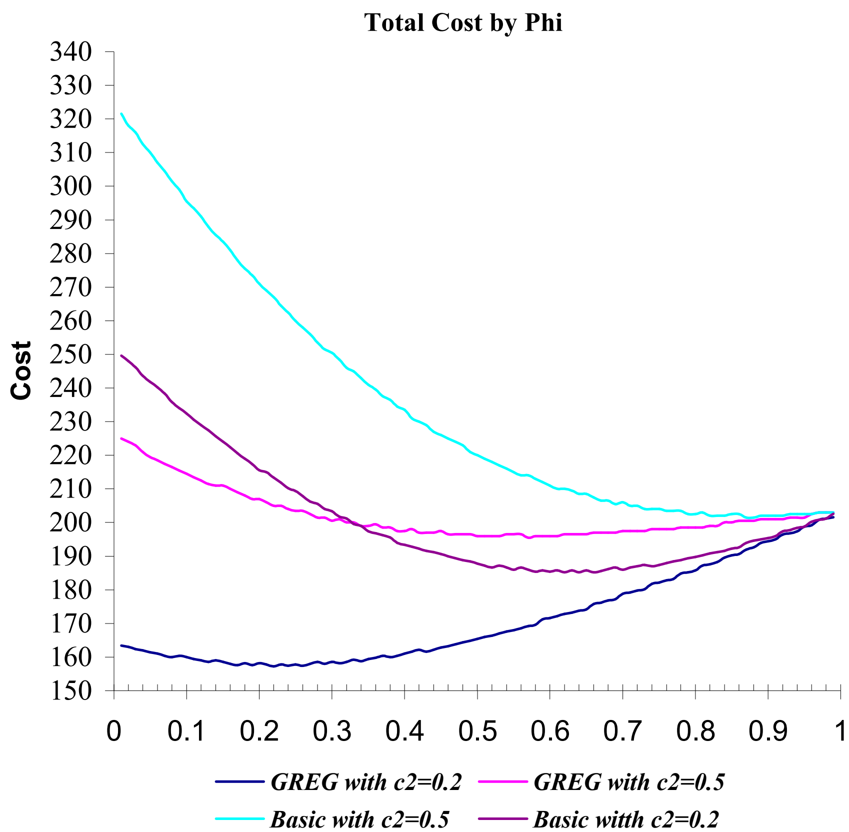

Figure 1: Minimum cost for different value of φ