# ON CALIBRATED ESTIMATOR FOR TWO-PHASE SAMPLING DESIGN IN THE PRESENCE OF NONRESPONSE

A. Demnati[1]

## ABSTRACT

Suppose a main sample is first selected for self-enumeration, and then from the main sample, we select in advance a subsample for follow-up. The final sample consists of respondents from the main sample and from the sub-sample. The design weights of the final sample are adjusted to compensate for nonresponse. We studied point and variance estimation under the above set-up. Results of a simulation study illustrate the effect of the weight adjustments on the mean square errors

KEY WORDS: Bernoulli sampling, Sample size determination, Unit nonresponse, Variance estimation.

## RÉSUMÉ

Supposons qu'un échantillon principal est d'abord sélectionné pour l'autodénombrement et qu'ensuite à partir de ce même échantillon, nous sélectionnons à l'avance un sous-échantillon pour le suivi. L'échantillon final se compose de répondants provenant de l'échantillon principal et du sous-échantillon. Les poids de sondage de l'échantillon final sont ajustés pour compenser la non-réponse. Nous étudions l'estimation ponctuelle et l'estimation de la variance en fonction du système ci-dessus. Les résultats de l'étude de simulation montrent les effets de la pondération sur les erreurs quadratiques moyennes

MOTS CLÉS : Échantillonnage de Bernoulli; détermination de la taille d'échantillon; non-réponse des unités ; estimation de la variance.

## 1. INTRODUCTION

Collecting information from sampled units by mail or over the Internet is much cheaper than through interviews, which makes self-enumeration an attractive data collection method for surveys. However, self-enumeration can produce high nonresponse rates in comparison to interviews. Currently, in some business surveys, in order to reduce the total cost of data collection, telephone follow-up for nonresponse is performed on only a portion of nonrespondents. These units are often identified in a deterministic way solely based, for example, on their expected contribution to the estimate of the total revenue. Clearly, this approach has the potential to introduce bias in the estimates. In addition, since a significant number of units are never followed up for nonresponse, the final unweighted nonresponse rate can be pretty high (50% in some cases). Unbiased estimates can be obtained by sub-sampling from nonrespondents (Hansen and Hurwitz, 1946). In reality, data collections from self-enumeration and from telephone follow-up are done in parallel, which makes sub-sampling from nonrespondents difficult to apply in some applications. The approach presented here would be a way to incorporate, using a two phase sampling framework, both the initial sample selection and the selection of units to follow-up for nonresponse. We consider a main sample for which values of the variables of interest are observed through self-enumeration, and an order subsample identified for follow-up activities. The order subsample is an alternative to a fixed size subsample when we are unsure of being able to complete all the follow-up required in the case of a fixed size subsample. In section 2, the proposed sampling scheme is described and a design-consistent estimator for the finite population total is studied. We determine the sample sizes that minimize the total cost subject to constraints on the total variance of estimators of the population totals under the sampling design and the nonresponse mechanism. Then we determine the order subsamples given the main sample. Weight adjustments to compensate for unit nonresponse in the subsample as well as to reduce the variance are studied in section 3.Variance estimation for a double calibrated estimator is given. Section 3 also presents the results of a limited simulation.

[1] Abdellatif Demnati, Business Survey Methods Division, Statistics Canada, Canada, K1A 0T6, Abdellatif.Demnati@statcan.gc.ca

## 2. THE SAMPLING SCHEME

### 2.1 Basic estimator of the population total

Suppose that a main sample for self-enumeration is selected from a finite population of size $N$, and let $d_{m;k} = a_{m;k} / \pi_{m;k}$ denote the associated design weights, where $a_{m;k}$ is the main sample membership indicator variable for element $k$, $\pi_{m;k} = E_p(a_{m;k})$, $E_p$ denotes expectation with respect to the sampling scheme, and the subscript $m$ in $d_{m;k}$ stands for "main". Let $r_{m;k}$ denote the main sample unit nonresponse indicator variable for enterprise $k$, i.e., $r_{m;k} = 0$ if there is unit nonresponse and $r_{m;k} = 1$ if there is partial response. From the main sample $s_m$, we select in advance a subsample $s$ of size $n$ for follow-up. Let $d_{2|m;k} = a_{2|m;k} / \pi_{2|m;k}$ denote the conditional subsample design weights where $a_{2|m;k}$ is the conditional subsample membership indicator variable and $\pi_{2|m;k} = E_p(a_{2|m;k} \mid a_{m;k} = 1)$. If all sub-sampled elements respond, then an estimator of the population total $Y = \sum_k y_k$ is given by

$$\hat{Y}^{(C)} = \sum_k d_{m;k} r_{m;k} y_k + \sum_k d_{f;k}(1 - r_{m;k}) y_k , \tag{2.1}$$

where $d_{f;k} = d_{m;k} d_{2|m;k}$, the subscript $f$ in $d_{f;k}$ stands for "follow-up", and the sum is over all the population units. The first part of the right term in (2.1) is an unbiased estimator of the total of the subpopulation using the self-enumeration method, and the second part is an unbiased estimator of the total of rest of the population. Hence, the estimator given by (2.1) is design unbiased for the entire population total, $E_p(\hat{Y}^{(C)}) = \sum_k r_{m;k} y_k + \sum_k (1 - r_{m;k}) y_k = \sum_k y_k$. We may write the estimator given by (2.1) as $\hat{Y}^{(C)} = \sum_k d_k y_k$, where $d_k = d_{m;k} r_{m;k} + d_{f;k}(1 - r_{m;k})$.

### 2.2 Variance

We consider first the estimator given by $\hat{U} = \sum_k d_{m;k} u_{m;k} + \sum_k d_{f;k} u_{f;k} = \sum_k \boldsymbol{u}_k^T \boldsymbol{d}_k$, where $\boldsymbol{d}_k = (d_{m;k}, d_{f;k})^T$, and $\boldsymbol{u}_k = (u_{m;k}, u_{f;k})^T$ is a vector of constants. The sampling variance, $Var_p(\boldsymbol{u})$, of the total $\hat{U}$ is given by

$$\begin{aligned} Var_p(\boldsymbol{u}) &= Var_p\left(\sum_k d_{m;k} u_{m;k}\right) + Var_p\left(\sum_k d_{f;k} u_{f;k}\right) + 2Cov_p\left(\sum_k d_{m;k} u_{m;k}, \sum_k d_{f;k} u_{f;k}\right) \\ &\equiv V_{m;p}(u_m) + V_{f;p}(u_f) + 2C_p(u_m, u_f). \end{aligned} \tag{2.2}$$

We consider stratified Bernoulli sampling (STBS) at both stages for a population having $H$ strata: a STBS of sampling fractions $\boldsymbol{f}_m = (f_{m;1}, ..., f_{m;h}, ..., f_{m;H})$ is obtained as the main sample; and then from each main sample stratum $s_{mh}$, a Bernoulli subsample $s_h$ of sampling fraction $f_{f;h}$ is selected. We assume the two vectors $\boldsymbol{f}_m$ and $\boldsymbol{f}_f$ fixed (Rao, 1973). Under the above sampling scheme, $V_{m;p}(u_m) = \sum_h \sum_k J_{hk}(f_{m;h}^{-1} - 1) u_{m;k}^2$, $V_{f;p}(u_f) = \sum_h \sum_k J_{hk}(f_h^{-1} - 1) u_{f;k}^2$, and $C_p(u_{m;}, u_f) = \sum_h \sum_k J_{hk}(f_{m;h}^{-1} - 1) u_{m;k} u_{f;k}$, where $f_h = f_{m;h} f_{f;h}$ and $J_{hk}$ is the stratum $h$ membership indicator for unit $k$.

We now turn to the derivation of the total variance of the linear combination $\hat{L} = \sum \{d_{m;k} r_{m;k} l_{m;k} + d_{f;k}(1 - r_{m;k}) l_{f;k}\}$ where $\boldsymbol{l}_k = (l_{m;k}, l_{f;k})^T$ is a vector of constants. We assume that the order of expectation can be interchanged so that $E_p E_{rm} = E_{rm} E_p$, where $E_{rm}$ denotes expectation with respect to the response mechanism associated with the main sample. We decompose the total variance of $\hat{L}$ as

$$Var_T(\hat{L}) = E_{rm} Var_p(\hat{L}) + Var_{rm} E_p(\hat{L}) \equiv V_{T1} + V_{T2}, \tag{2.3}$$

where $Var_{rm}$ denotes the variance with respect to the response mechanism associated with the main sample. The first term $V_{T1} = E_{rm} Var_p(\hat{L})$ of (2.3) is given by

$$V_{T1} = -\sum_k \{p_{m;k} l_{m;k}^2 + (1 - p_{m;k}) l_{f;k}^2\} + \sum_h \sum_k J_{hk} p_{m;k} l_{m;k}^2 / f_{m;h} + \sum_h \sum_k J_{hk}(1 - p_{m;k}) l_{f;h}^2 / f_h , \tag{2.4}$$

where $p_{m;k} = E_{rm}(r_{m;k})$. Under an independent response mechanism, the second term $V_{T2} = Var_{rm} E_p(\hat{L})$ of (2.3) is given by

$$V_{T2} = \sum_k p_{m;k}(1 - p_{m;k})(l_{m;k} - l_{f;k})^2 . \tag{2.5}$$

It follows from (2.3), (2.4), and (2.5) that, under STBS at both stages, we can express $Var_T(\hat{L})$ as

$$Var_T(\hat{L}) = v_0 + \sum_h v_{mh} / f_{m;h} + \sum_h v_h / f_h , \tag{2.6}$$

where $v_0 = \sum_k p_{m;k}(1-p_{m;k})(l_{m;k} - l_{f;k})^2 - \sum_k \{p_{m;k} l_{m;k}^2 + (1-p_{m;k}) l_{f;k}^2\}$, $v_{mh} = \sum_k J_{hk} p_{m;k} l_{m;k}^2$, and $v_h = \sum_k J_{hk}(1-p_{m;k}) l_{f;k}^2$.

### 2.3 Sample size determination

We assume that obtaining values through self-enumeration is much cheaper than obtaining values through follow-up. Consider $I$ population characteristics of interest denoted by $y_1,...,y_I$, and we want to determine $\boldsymbol{f}_m$ and $\boldsymbol{f}_f$ for estimating the corresponding finite population parameters $Y_i$ $(i = 1,...,I)$ which minimize the cost subject to constraints on variances. It follows from (2.6), that, under STBS at both stages, $Var_T(\hat{Y}_i)$, for the $i^{th}$ variable $y_i$, reduces to

$$Var(\hat{Y}_i^{(C)}) = v_{i0} + \sum_h v_{imh} / f_{m;h} + \sum_h v_{ih} / f_h, \; i = 1,...,I, \tag{2.7}$$

where $v_{i0} = -\sum y_{ik}^2$, $v_{imh} = \sum_k J_{hk} p_{m;k} y_{ik}^2$, and $v_{ih} = \sum_k J_{hk}(1-p_{m;k}) y_{ik}^2$.

We determine $f_{m;h}$ and $f_{f;h}$ such that the cost

$$C = c_0 + \sum_h \{C_{mh} f_{m;h} + \sum_h C_h f_h\},$$

is minimized subject to constraints on the variance

$$Var(\hat{Y}_i^{(C)}) \le V_i, i = 1,...,I,$$

and constraints on sampling fractions

$$0 < f_{m;h} \le 1, \; 0 < f_{f;h} \le 1, \; h = 1,...,H.$$

where $c_0$ is the overhead cost, $C_{mh} = (c_{hq} + c_{hp}\bar{p}_{m;h}) N_h$, $c_{hq}$ is the cost per element of sending the questionnaire in stratum $h$, $c_{hp}$ is the cost per element of processing the received questionnaire, $\bar{p}_{m;h} = N_h^{-1} \sum J_{hk} p_{m;k}$ is the expected number of completed questionnaires through self-enumeration in stratum $h$, $C_h = (c_{hp} + c_{hf})(1-\bar{p}_{m;h}) N_h$, $c_{hf}$ is the cost per element for follow-up in stratum $h$ and the $V_i$ are specified tolerances. For example, one could specify an upper limit, $\tau_i$, on the coefficient of variation of $\hat{Y}_i^{(C)}$ so that $V_i = (\tau_i Y_i)^2$. In the case of complete response ( $p_{m;k} = 1$), the estimator $\hat{Y}_i^{(C)}$ reduces to the HT estimator $\sum_k d_{m;k} y_{ik}$ with $C = c_0 + \sum_h C_{mh} f_{m;h}$, $C_{mh} = (c_{hq} + c_{hp}) N_h$, $C_h = 0$, $v_{i0} = -\sum_k y_{ik}^2$, $v_{imh} = \sum_k J_{hk} y_{ik}^2$, and $v_{ih} = 0$.

### 2.4. Order subsample for follow-up

In subsection 2.1, we succeeded in eliminating the potential bias coming from the nonresponse occurring in the non followed up portion of the main sample. Let us now focus on the subsample selected for follow-up. Since predicting in advance how well data collection will progress is not always easy, it may very well happen that not all non responding units in the subsample end up being followed up, or on the opposite, it is also possible that the capacity to follow up non responding units ends up being more than expected. Since our objective is to be as close as possible to 100% response in the subsample (in order to use the unbiased estimator presented in subsection 2.1), while at the same time making an efficient use of the collection resources at our disposal, we propose an order subsample as an alternative to the fixed size subsample.

The method would basically consist of selecting a nested set of subsamples and a set of stopping rules. After the follow up activities in a certain group are complete the stopping rules are consulted to see if the active collection period should stop. If not, the interviewers move on to the next group of units. We do not, therefore, use fixed subsample size, although, an expected subsample size is always specified in the sampling plan. The groups of elements are formed by dividing the main sample into nested subsamples and assigning an order for interviews to each group based on the coefficients of variations of the estimates. Using STBS, we proceed as follows:

a) Assign a unique random number $r_k$, generated from the uniform distribution, to each enterprise in the main sample.

b) Given the main-sample with sampling fractions $\boldsymbol{f}_m$, determine sampling fraction $\boldsymbol{f}_f^{(t)} = (f_{f;1}^{(t)},..., f_{f;H}^{(t)})$ of the order subsample $t$ by minimizing the survey cost subject to constraints on the variances using an upper limit $\tau^{(t)}$ on the coefficients of variation, i.e., we determine $f_{f;h}^{(t)}$ $(h = 1,...,H)$ such that the cost

$$C = c_0^* + \sum_h C_h^* f_{f;h}^{(t)},$$

is minimized subject to constraints on the variance

$$Var(\hat{Y}_i^{(C)}) = v_{i0}^* + \sum_h v_{ih}^* / f_{f;h}^{(t)} \le V_i^{(t)},\ i = 1,...,I,$$

and constraints on sampling fractions

$$f_{f;h}^{(t-1)} \le f_{f;h}^{(t)} \le 1,\ h = 1,...,H,$$

with $f_{f;h}^{(0)} = 0$, where $c_0^* = c_0 + \sum_h c_{m;h} f_{m;h}$, $C_h^* = C_{mh} f_{m;h}$, $v_{i0}^* = v_{i0} + \sum_h v_{imh} / f_{m;h}$, and $v_{ih}^* = v_{ih} / f_{m;h}$. The quantities $v_{i0}$, $v_{imh}$, and $v_{ih}$ are defined in subsection 2.3.

c) Form the order subsample $s_h^{(t)}$ with element of $s_{m;h}$ having $r_k \le f_{f;h}^{(t)}$

d) Repeat steps b) and c) for $t = 1,...,T$ with $\tau^{(1)} > \tau^{(2)} > ... > \tau^{(T)}$ respectively until the desired $\tau$.

We have $s_h^{(1)} \subseteq s_h^{(2)} \subseteq .... \subseteq s_h^{(T)}$. Elements are followed-up in increasing order of sub-sampling: elements of the first group, $s_{h;1} = s_h^{(1)} \setminus \varnothing$, are first followed-up, followed by elements of the second group, $s_{h;2} = s_h^{(2)} \setminus s_h^{(1)}$, and so on, where $\varnothing$ denotes the empty set, $A \setminus B = A \cap B^c$ and $B^c$ is the complement of the set $B$ with respect to $A \cup B$. Depending on the stopping rules, only a few groups may be sufficient to stop the collection process. If we stop collection after following up the group $s_{h;t} = s_h^{(t)} \setminus s_h^{(t-1)}$ then the final subsample is set to $s_h = s_h^{(t)}$. So, we can discard non responding units of groups $s_{h,l}$ for $l = t+1,...,T$, without creating any bias. The final subsample $s = \cup s_h$ is a STBS with sampling fractions $f_{f;h} = f_{f;h}^{(t)},\ h = 1,...,H$.

## 3. UNIT NONRESPONSE IN THE SUBSAMPLE

### 3.1 Weight adjustments for unit nonresponse and for variance reduction

Up until now, it was assumed that 100% response would be achieved within the subsample. Obviously, this assumption will most likely never be fulfilled and therefore, it is more realistic to assume a certain level of nonresponse in the subsample. A first widely-used weight adjustment approach to compensate for unit nonresponse is to define a new set of weights, $\tilde{d}_{f;k}^{(1)}$ with $k^{th}$ element equals to

$$\tilde{d}_{f;k}^{(1)} = d_{f;k}(1 - r_{m;k}) r_{f;k} / \hat{p}_{f;k}, \tag{3.1}$$

where $r_{f;k}$ is the subsample response indicator variable for element $k$. We assume that the model on the response indicators $r_{f;k}$ is specified by a logistic regression model with mean $E_{rf}(r_{f;k}) = p_{f;k}(\boldsymbol{x}_k^T \boldsymbol{\alpha}) \equiv p_{f;k}$ where $\boldsymbol{x}_k$ is the vector of explanatory variables, $\boldsymbol{\alpha}$ is the model vector parameter and $E_{rf}$ denotes expectation with respect to the response mechanism related to the follow-up subsample. Here, $\hat{p}_{f;k} = p_{f;k}(\boldsymbol{x}_k^T \hat{\boldsymbol{\alpha}})$ and the vector $\hat{\boldsymbol{\alpha}}$ is the solution to the set of estimating equations $\hat{\boldsymbol{U}}(\boldsymbol{\alpha}) = \sum_k d_{f;k}(1 - r_{m;k}) \boldsymbol{x}_k (r_{f;k} - p_{f;k})$ with $p_{f;k} = \exp(\boldsymbol{x}_k^T \boldsymbol{\alpha}) / \{1 + \exp(\boldsymbol{x}_k^T \boldsymbol{\alpha})\}$. A common approach to handle unit nonresponse is to classify respondents and nonrespondents into $q$ adjustment classes, using auxiliary information on all sample element in which case $x_{ck}$ denotes the class $c$, $c = 1,...,q$, membership indicator variable for element $k$ with $\sum_c x_{ck} = 1$. In this case, if element $k$ belongs to class $c$ then $\hat{p}_{f;k} = \hat{N}_{cfr} / \hat{N}_{cf}$, $c = 1,...,q$, where $\hat{N}_{cf} = \sum_k d_{f;k}(1 - r_{m;k}) x_{ck}$ is the estimate of the size of class $c$ and $\hat{N}_{cfr} = \sum_k d_{f;k}(1 - r_{m;k}) r_{f;k} x_{ck}$ is the estimate of the number of respondents in class $c$.

A second widely-used weight adjustment approach to compensate for unit nonresponse is to define a new set of weights, $\tilde{d}_{f;k}^{(2)}$ with $k^{th}$ element equal to

$$\tilde{d}_{f;k}^{(2)} = d_{f;k}(1 - r_{m;k}) r_{f;k} g_{f;k}, \tag{3.2}$$

with $g_{f;k} = 1 + (\hat{\boldsymbol{X}}_f - \hat{\boldsymbol{X}}_{fr})^T [\sum_k d_{f;k}(1 - r_{m;k}) r_{f;k} \boldsymbol{x}_k \boldsymbol{x}_k^T]^{-1} \boldsymbol{x}_k$, where $\hat{\boldsymbol{X}}_{fr} = \sum_k d_{f;k}(1 - r_{m;k}) r_{f;k} \boldsymbol{x}_k$, $\hat{\boldsymbol{X}}_f = \sum_k d_{f;k}(1 - r_{m;k}) \boldsymbol{x}_k$, and $\boldsymbol{x}_k$ is the vector of calibration variables. The calibration weights, $\tilde{d}_{f;k}^{(2)}$ have the calibration properties: $\sum_k \tilde{d}_{f;k}^{(2)} \boldsymbol{x}_k = \hat{\boldsymbol{X}}_f$. The factor $g_{f;k}$ reduces to $\hat{p}_{f;k}$, when the calibration variables denote the class membership indicators.

Suppose an additional vector of calibration variables $\boldsymbol{t}_k$ with known totals $\boldsymbol{T}$ is available in addition to the vector $\boldsymbol{x}_k$. In this case the final weights are of the form $w_k = \tilde{d}_k g_k$, with $g_k = 1+(\boldsymbol{T}-\hat{\boldsymbol{T}})^T[\sum_k \tilde{d}_k \boldsymbol{t}_k \boldsymbol{t}_k^T]^{-1}\boldsymbol{t}_k$, $\tilde{d}_k = d_{m;k} r_{m;k} + \tilde{d}_{f;k}$ and $\tilde{d}_{f;k}$ denotes either $\tilde{d}_{f;k}^{(1)}$ or $\tilde{d}_{f;k}^{(2)}$. The double calibrated estimator of the domain total $Y_i$ is given by.

$$\tilde{\tilde{Y}}_i = \sum_k w_k y_{ik}. \tag{3.3}$$

### 3.2 Variance estimation

We write $\tilde{\tilde{Y}}_i$ as $f_i(\boldsymbol{A}_d)$, where $\boldsymbol{A}_d$ is $3\times N$ matrix with $k^{th}$ column $\boldsymbol{d}_k = (d_{1k}, d_{2k}, d_{3k})^T$, $d_{1k} = d_{m;k} r_{m;k}$, $d_{2k} = d_{f;k}(1-r_{m;k})$ and $d_{3k} = d_{f;k}(1-r_{m;k}) r_{f;k}$. Let $E$ be the total expectation, and $\boldsymbol{\mu}_k = E(\boldsymbol{d}_k)$. We assume that $f_i(\boldsymbol{A}_\mu) = \theta_i$, where $\boldsymbol{A}_\mu$ is a $3\times N$ matrix with $k^{th}$ column $\boldsymbol{\mu}_k$. Following Demnati and Rao (2010), a sampling variance estimator, $\vartheta(\tilde{\tilde{Y}}_i)$, of $\tilde{\tilde{Y}}_i$ is simply given by $\vartheta(\tilde{\tilde{Y}}_i) = \vartheta_T(z_i)$, with $\boldsymbol{z}_{ik} = \partial f_i(\boldsymbol{A}_b)/\partial \boldsymbol{b}_k \,|_{\boldsymbol{A}_b = \boldsymbol{A}_d}$, where $\boldsymbol{A}_b$ is the $3\times N$ matrix of arbitrary real numbers with $k^{th}$ column $\boldsymbol{b}_k = (b_{1k}, b_{2k}, b_{3k})^T$. Taking the derivatives, we get

$$\boldsymbol{z}_{ik}^{(1)} = \begin{cases} e_k \\ \boldsymbol{x}_k^T \hat{\boldsymbol{B}}(e)\; p(\boldsymbol{x}_k^T\hat{\boldsymbol{\alpha}}) \\ \dfrac{1}{p(\boldsymbol{x}_k^T\hat{\boldsymbol{\alpha}})}\left\{e_k - \boldsymbol{x}_k^T\hat{\boldsymbol{B}}(e)\, p(\boldsymbol{x}_k^T\hat{\boldsymbol{\alpha}})\right\} \end{cases} \quad \text{or} \quad \boldsymbol{z}_{ik}^{(2)} = \begin{cases} e_k \\ \boldsymbol{x}_k^T\hat{\boldsymbol{A}}(e) \\ g_{f;k}(e_k - \boldsymbol{x}_k^T\hat{\boldsymbol{A}}(e)) \end{cases}$$

where $\hat{\boldsymbol{B}}(u) = [\hat{\boldsymbol{J}}(\hat{\boldsymbol{\alpha}})]^{-1}\sum_k \{d_{3k}/p(\boldsymbol{x}_k^T\hat{\boldsymbol{\alpha}})\}\{1-p(\boldsymbol{x}_k^T\hat{\boldsymbol{\alpha}})\}\boldsymbol{x}_k u_k$, $\hat{\boldsymbol{J}}(\boldsymbol{\alpha}) = \sum_k d_{2k} p(\boldsymbol{x}_k^T\boldsymbol{\alpha})\{1-p(\boldsymbol{x}_k^T\boldsymbol{\alpha})\}\boldsymbol{x}_k\boldsymbol{x}_k^T$,
$e_k = g_k(y_k - \boldsymbol{t}_k^T\hat{\boldsymbol{C}}(y))$, $\hat{\boldsymbol{C}}(y) = \{\sum \tilde{d}_k \boldsymbol{t}_k\boldsymbol{t}_k^T\}^{-1}\sum \tilde{d}_k \boldsymbol{t}_k y_k$, and $\hat{\boldsymbol{A}}(u) = \{\sum d_{3k}\boldsymbol{x}_k\boldsymbol{x}_k^T\}^{-1}\sum d_{3k}\boldsymbol{x}_k u_k$.

We use $\boldsymbol{z}_{ik}^{(1)}$ when $d_{f;k}^{(1)}$ is used and $\boldsymbol{z}_{ik}^{(2)}$ when $d_{f;k}^{(2)}$ is used. It remains to derive $\vartheta_T(z_i)$.

We consider first the estimation of the sampling variance of the linear combination $\hat{U} = \sum_k \boldsymbol{u}_k^T\boldsymbol{d}_k$ where $\boldsymbol{d}_k = (d_{m;k}, d_{f;k})^T$ and $\boldsymbol{u}_k = (u_{m;k}, u_{f;k})^T$ a vector of constants. An estimator, $\vartheta_p(\boldsymbol{u})$, of the sampling variance of $\hat{U}$ is given by

$$\vartheta_p(\boldsymbol{u}) = \vartheta_{m;p}(u_m) + \vartheta_{f;p}(u_f) + 2c_p(u_m, u_f), \tag{3.4}$$

where $\vartheta_{m;p}(u_{m;k}) = \sum_k d_{m;k}(\pi_{m;k}^{-1}-1)u_{m;k}^2$, $\vartheta_{f;p}(u_{f;k}) = \sum_k d_{f;k}(\pi_{f;k}^{-1}-1)u_{f;k}^2$, and $c_p(u_m, u_f) = \sum_k d_{f;k}(\pi_{m;k}^{-1}-1)u_{m;k}u_{f;k}$.

We now turn to the estimation of the total variance of the linear combination $\hat{L} = \sum_k \boldsymbol{l}_k^T\boldsymbol{d}_k$, where $\boldsymbol{d}_k = (d_{1k}, d_{2k}, d_{3k})^T$, $\boldsymbol{l}_k = (l_{1k}, l_{2k}, l_{3k})^T$ a vector of constants, $d_{1k} = d_{m;k}r_{m;k}$, $d_{2k} = d_{f;k}(1-r_{m;k})$ and $d_{3k} = d_{f;k}(1-r_{m;k})r_{f;k}$. We decompose the total variance of $\hat{L}$ as

$$Var_T(\hat{L}) = E_{rf}E_{rm}Var_p(\hat{L}) + E_{rf}Var_{rm}E_p(\hat{L}) + Var_{rf}E_{rm}E_p(\hat{L}) = V_p + V_{rm} + V_{rf}, \tag{3.5}$$

where $Var_{rf}$ denotes variance with respect to the response mechanism related to the subsample.

An estimator of the first component $V_p$ of the total variance given by (3.5) is given by (3.4) with $u_{m;k} = l_{1k}r_{m;k}$ and $u_{f;k} = (1-r_{m;k})\{l_{2k} + l_{3k}r_{f;k}\}$.

An estimator of the second component $V_{rm}$ of the total variance given in (3.5) is obtained by estimating $Var_{rm}\sum_k\{v_{m;k}r_{m;k} + v_{f;k}(1-r_{m;k})\}$, where $v_{m;k} = l_{1k}$ and $v_{f;k} = l_{2k} + l_{3k}r_{f;k}$. The formula of a linear combination of random variables gives

$$\begin{aligned} Var_{rm}\{\textstyle\sum_k v_{m;k}r_{m;k} + \sum_k v_{f;k}(1-r_{m;k})\} &= Var_{rm}\{\textstyle\sum_k v_{m;k}r_{m;k}\} \\ &\quad + Var_{rm}\{\textstyle\sum_k v_{f;k}(1-r_{m;k})\} \\ &\quad + 2Cov_{rm}\{\textstyle\sum_k v_{m;k}r_{m;k}, \sum_k v_{f;k}(1-r_{m;k})\}. \end{aligned}$$

An estimator of $Var_{rm}\{\sum_k v_{m;k} r_{m;k}\}$ is given by $\sum_k d_{m;k} r_{m;k}(1-\hat{p}_{m;k}) v_{m;k}^2$, and an estimator of $Var_{rm}\{\sum_k v_{f;k}(1-r_{m;k})\}$ is given by $\sum_k d_{f;k} \hat{p}_{m;k}(1-r_{m;k}) v_{f;k}^2$, where $\hat{p}_{f;k} = \hat{E}_{rm}(r_{m;k})$ denote an estimator of $p_{f;k} = E_{rf}(r_{f;k})$. Now $Cov_{rm}\{\sum_k v_{m;k} r_{m;k}, \sum_k v_{f;k}(1-r_{m;k})\} = -\sum_k v_{m;k} v_{f;k} p_{m;k}(1-p_{m;k})$. If $v_{m;k}$ is observable in the subsample then an estimator of $Cov_{rm}\{\sum_k v_{m;k} r_{m;k}, \sum_k v_{f;k}(1-r_{m;k})\}$ is given by $-\sum_k d_{f;k} v_{m;k} v_{f;k} \hat{p}_{m;k}(1-r_{m;k})$.

An estimator of the third component $V_{rf}$ of the total variance given by (3.5) is given by an estimator of $Var_{rf}(\sum_k (1-p_{m;k}) l_{3k} r_{f;k})$. Under an independent response mechanism, an estimator of $V_{rf}$ is given by

$$\hat{V}_{rf} = \sum_k d_{f;k}(1-r_{m;k}) r_{f;k}(1-\hat{p}_{m;k})(1-\hat{p}_{f;k}) l_{3k}^2 .$$

An estimator of the total variance of $\hat{L}$ is given by the sum of estimators of the three components.

### 3.3 Simulation

We generate one finite population of size $N = 5000$ from the model $y_k = x_k + 2x_k^{1/2}\varepsilon_k$, using $x_k = 50 + 10u_k$ where $\varepsilon_k$ and $u_k$ are independent observations from $N(0,1)$. We stratified the population into two strata with 2085 units having $x_k < 48$ in stratum 1 and 2915 units having $x_k \geq 48$. We selected $B = 2000$ stratified two-phase Bernoulli samples with sampling fractions: $f_{m;1} = .4$, $f_{m;2} = .25$ for the main-sample and $f_{f;1} = .2$, $f_{f;2} = .5$ for the conditional subsample. We generated the probabilities of response from $logit(p_{m;k}) = -.01 x_k$ and $logit(p_{f;k}) = .03 x_k$. These choices give an average response rate of about 38% for the main sample and 81% for the follow-up. We used $x_k$ for nonresponse adjustment through logistic regression, and set $\boldsymbol{x}_k = (J_{1k}, J_{2k})^T$ for the GREG nonresponse adjustment. We then set $\boldsymbol{t}_k = (J_{1k}, J_{2k})^T$ for the final adjustment. The parameter of interest is the population total $\theta_N = Y$. Let $\hat{\theta}$ denote an estimator of the population total, and $\vartheta(\hat{\theta})$ be the associated total variance estimator. We calculated $\hat{\theta}$ and associated total variance estimate $\vartheta(\hat{\theta})$, from each simulated sample $b$ ($b = 1,..., B$) and their averages $\bar{\hat{\theta}}$ and $\bar{\vartheta}(\hat{\theta})$ over $b$. The simulated total mean squared error (MSE) of estimator $\hat{\theta}$ is calculated as $M(\hat{\theta}) = B^{-1}\sum_{b=1}^{B}(\hat{\theta}_b - \theta_N)^2$, where $\hat{\theta}_b$ is the value of $\hat{\theta}$ for the $b^{th}$ simulated sample. The simulated relative bias of $\hat{\theta}$ and $\vartheta(\hat{\theta})$ are calculated as $RB(\hat{\theta}) = (\bar{\hat{\theta}} - \theta_N)/\theta_N$ and $RB(\vartheta(\hat{\theta})) = (\bar{\vartheta}(\hat{\theta}) - M(\hat{\theta}))/M(\hat{\theta})$, and shown in Table 1. The final column of Table 1 compares the MSE to the MSE for the estimator $\hat{Y}$ using only nonresponse adjustment through logistic regression. It is seen that both adjustments for nonresponse are efficient. The relative error of the total variance estimator is negligible. Post-stratification to strata sizes is highly efficient. The reduction in the MSE is around 90%.

Table 1: Simulation results for the double calibrated estimator of the population total

| | | $RB(\hat{\theta})$ | $RB(\vartheta(\hat{\theta}))$ | $M(\hat{\theta})/M(\hat{Y})$ |
|---|---|---|---|---|
| Adjustment for nonresponse | Logistic | .0005 | .009 | .09 |
| | GREG | .003 | -.04 | .10 |

## REFERENCES


Demnati, A. and Rao, J.N.K. (2010). Linearization variance estimators for model parameters from complex survey data. *Survey Methodology*, **36**, 193-199.

Hansen, M.H. and Hurwitz, W.N. (1946). The problem of nonresponse in sample survey, *Journal of the American Statistical Association*, **41**, 517-529.

Rao, J.N.K. (1973). On double sampling for stratification and analytic surveys. *Biometrika*, **60**, 125-133.