Linearization Variance Estimation and Allocation for Two-phase Sampling under Mass Imputation

A. Demnati and J.N.K. Rao

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

Carleton University
School of Mathematics and Statistics
Ottawa, Canada, K1S 5B6
JRao@math.carleton.ca

Abstract

We consider two-phase sampling in which values of a variable of interest are observed only in the second-phase sub-sample. Values for the first-phase units not sampled in the second-phase are mass imputed, using values from an administrative file when available and regression imputation otherwise. Such two-phase sampling methods are often used in annual business surveys to reduce survey costs and respondent burden, assuming that collecting values from administrative sources is much cheaper than obtaining values through questionnaires from sampled units. We study both naïve and design-consistent estimators for a total or mean under the above set-up. We also obtain associated variance estimators using a unified approach proposed by Demnati and Rao (2004). Simulation results on the finite sample performance of the estimators and associated variance estimators are also presented, using substitution or ratio mass imputation. We also study the case of missing sub-sample values and develop estimators of the total and associated variance estimators. Sample allocation issues are also studied.

KEY WORDS: Regression imputation; Sample sizes determination; Substitution method; Two-phase sampling.

1. Introduction

At Statistics Canada, more than sixty annual business surveys are carried out using a holistic design termed as the Unified Enterprise Survey (UES). One of the objectives of the UES is to reduce survey cost and response burden by replacing some questionnaires with values from administrative data. We assume that obtaining values from administrative data is much cheaper than obtaining values through questionnaires. A random sample of size m units is first selected. Then a sub-sample of size n is selected and surveyed. The remaining $m-n$ units as well as non-respondents are imputed using administrative files. Composite imputation involving two or more different methods is also often used; for example, the values from administrative file (e.g., tax file), when available, and regression imputation otherwise.

Demnati and Rao (2008) considered the case of simulated census data generated from a probability sample by imputing for the non-sampled units and sample non-respondents using auxiliary variables. They studied the estimation of a finite population total and other parameters. In this paper, we extended their work to two-phase sampling. For variance estimation, we have used the unified approach proposed by Demnati and Rao (2004). We have studied naïve estimators based on the simulated first-phase data as well as design-consistent estimators of a population total. In section 2, we consider the case of complete response and report the results of a simulation study. The case of missing data is studied in section 3. In section 4, we consider stratified simple random sampling in both phases and obtain "optimal" first-phase and second-phase strata sample sizes, m_h and n_h, that minimize the cost subject to constraints on the variances of estimators for one or more characteristics of interest.

2. Complete Response

2.1 Imputed Estimators

We first consider the case where complete responses are obtained from the second-phase subsample. Imputation is performed on all the first-phase units not sampled in the second-phase. Here, the imputed values $\hat{y}_k^*$ are given by $\hat{y}_k^* = (1-I_k)\hat{y}_{1k}^* + I_k\hat{y}_{2k}^*$ with $\hat{y}_{1k}^* = t_k$ and $\hat{y}_{2k}^*$ based on an imputation model, where t_k is the value from an administrative file and the constant I_k is the missing t_k indicator.

A naïve estimator of the finite population total $Y = \sum y_k$ is given by

$$\hat{Y}^{(N)} = \sum d_k^{(1)}\{a_k^{(2|1)} y_k + (1-a_k^{(2|1)})\hat{y}_k^*\}, \tag{2.1}$$

where $d_k^{(1)} = a_k^{(1)}/\pi_k^{(1)}$, $a_k^{(1)}$ is the first-phase sample membership indicator variable, $\pi_k^{(1)} = E_p(a_k^{(1)})$, E_p denotes design expectation, and $a_k^{(2|1)}$ is the conditional second-phase sample membership indicator variable.

Suppose the imputation model on the responses y_k is specified by a generalized linear model with mean $E_m(y_k) = \mu_k(\boldsymbol{\beta}) = h(\boldsymbol{x}_k^T\boldsymbol{\beta})$, where $\boldsymbol{x}_k$ is a $p\times 1$ vector of explanatory variables for the model mean, $h()$ is a "link" function and E_m denotes model expectation. For example, the choice $h(a) = a$ gives a linear regression model, and $h(a) = e^a/(1+e^a)$ gives a logistic regression model for binary responses y_k. The imputed values $\hat{y}_{2k}^*$ are given by $\hat{y}_{2k}^* = \mu_k(\hat{\boldsymbol{\beta}}^{(N)})$, where the naïve estimator $\hat{\boldsymbol{\beta}}^{(N)}$ is obtained as the solution to the estimating equations

$$\hat{\boldsymbol{l}}_\beta^{(N)}(\boldsymbol{\beta}) = \sum a_k I_k \boldsymbol{\Phi}_k(\boldsymbol{\beta})(y_k - \mu_k(\boldsymbol{\beta})) = 0, \tag{2.2}$$

$a_k = a_k^{(1)}a_k^{(2|1)}$ and $\boldsymbol{\Phi}_k(\boldsymbol{\beta}) = [\partial\mu_k(\boldsymbol{\beta})/\partial\boldsymbol{\beta}][Var_m(y_k)]^{-1}$. For example, $\boldsymbol{\Phi}_k(\boldsymbol{\beta}) = \boldsymbol{x}_k$ for linear and logistic regression models.

The finite population parameter θ_N induced by the estimator $\hat{Y}^{(N)}$ is given by

$$\theta_N = E_p(\hat{Y}^{(N)}) \approx \sum \pi_k^{(2|1)} y_k + \sum(1-\pi_k^{(2|1)})y_k^*, \tag{2.3}$$

where $y_k^* = E_p(\hat{y}_k^*)$. It is clear from (2.3) that θ_N depend on the selection probabilities. For example, the use of $\hat{Y}^{(N)}$ under simple random sampling at both stages induces a mixture of two totals as finite population parameter: $\theta_N \approx f^{(2|1)}\sum y_k + (1-f^{(2|1)})\sum y_k^*$ where $f^{(2|1)}$ is the conditional sampling fraction for the second-phase sample.

The sampling bias induced by the estimator $\hat{Y}^{(N)}$ in estimating the finite population total Y is given by

$$\theta_N - Y \approx -\sum(1-\pi_k^{(2|1)})(y_k - y_k^*) \equiv B. \tag{2.4}$$

In order to remove the sampling bias, one may first estimate the bias (2.4) by

$$\hat{B} = -\sum d_k(1-\pi_k^{(2|1)})(y_k - \hat{y}_k^*), \tag{2.5}$$

and then adjust $\hat{Y}^{(N)}$ to get the bias-adjusted estimator

$$\hat{Y} = \sum d_k y_k + \sum(d_k^{(1)} - d_k)\hat{y}_k^*, \tag{2.6}$$

with $\hat{y}_{2k}^* = \mu_k(\hat{\boldsymbol{\beta}})$, the estimator $\hat{\boldsymbol{\beta}}$ is obtained as solution to the estimating equations

$$\hat{\boldsymbol{l}}_\beta(\boldsymbol{\beta}) = \sum d_k I_k \boldsymbol{\Phi}_k(\boldsymbol{\beta})(y_k - \mu_k(\boldsymbol{\beta})) = 0, \tag{2.7}$$

where $d_k = a_k/\pi_k$ and $\pi_k = E_p(a_k)$. The design-based estimator $\hat{Y}$ is approximately unbiased for Y: $E_p(\hat{Y}) \approx Y$. We may also write $\hat{Y}$ as

$$\hat{Y} = \sum d_k^{(1)}\{d_k^{(2|1)} y_k + (1-d_k^{(2|1)})\hat{y}_k^*\}.$$

2.2 Variance Estimation

We suppose first that the parameter of interest is $\theta_N = E_p(\hat{Y}^{(N)})$. Let $\boldsymbol{d}_k = (d_{1k}, d_{2k})^T$, where $d_{1k} = d_k^{(1)}$, and $d_{2k} = d_k$. The Demnati–Rao (DR) variance estimator (Demnati and Rao, 2004) is simply given by

$$\vartheta_{DR}(\hat{Y}^{(N)}) = \vartheta(\boldsymbol{z}), \tag{2.8}$$

where $\vartheta(\boldsymbol{u})$, in operation notation, is the variance estimator of the linear combination $\hat{U} = \sum \boldsymbol{u}_k^T \boldsymbol{d}_k$, $\boldsymbol{u}_k = (u_{1k}, u_{2k})^T$ a vector of constants, $\boldsymbol{z}_k = \partial f(\boldsymbol{A}_b)/\partial \boldsymbol{b}_k \mid_{\boldsymbol{A}_b = \boldsymbol{A}_d}$, $f(\boldsymbol{A}_d) = \hat{Y}^{(N)}$, $\boldsymbol{A}_d$ is a $2 \times N$ matrix with k^{th} column $\boldsymbol{d}_k$, and $\boldsymbol{A}_b$ is a $2 \times N$ matrix of arbitrary real numbers with k^{th} column $\boldsymbol{b}_k = (b_k^{(1)}, b_k)^T$. The design based variance estimator $\vartheta(\boldsymbol{u})$ of the total $\hat{U}$ is given by

$$\vartheta(\boldsymbol{u}) = \sum_k \sum_t \boldsymbol{u}_k^T \operatorname{cov}_p(\boldsymbol{d}_k, \boldsymbol{d}_t) \boldsymbol{u}_t, \tag{2.9}$$

with

$$\operatorname{cov}_p(\boldsymbol{d}_k, \boldsymbol{d}_t) = diag(\boldsymbol{d}_k) \begin{pmatrix} 1-\omega_{kt}^{(1)} & 1-\omega_{kt}^{(1)} \\ 1-\omega_{kt}^{(1)} & 1-\omega_{kt} \end{pmatrix} diag(\boldsymbol{d}_t), \tag{2.10}$$

where $\omega_{kt}^{(1)} = \pi_k^{(1)} \pi_t^{(1)} / \pi_{kt}^{(1)}$, $\omega_{kt} = \pi_k \pi_t / \pi_{kt}$, and $(\pi_{kt}^{(1)}, \pi_{kt})$ denote respectively the first and two phase joint inclusion probabilities. Substituting (2.10) into (2.9) we get

$$\vartheta(\boldsymbol{u}) = \vartheta_s^{(1)}(u_1, u_1) + 2c_s(u_1, u_2) + \vartheta_s(u_2, u_2), \tag{2.11}$$

with

$$\vartheta_s^{(1)}(x, y) = \sum\sum d_k^{(1)} d_t^{(1)} (1-\omega_{kt}^{(1)}) x_k y_t, \tag{2.12}$$

$$c_s(x, y) = \sum\sum d_k^{(1)} d_t (1-\omega_{kt}^{(1)}) x_k y_t, \tag{2.13}$$

and

$$\vartheta_s(x, y) = \sum\sum d_k d_t (1-\omega_{kt}) x_k y_t. \tag{2.14}$$

We may write $c_s(x, y)$ as

$$c_s(x, y) = \sum\sum d_k^{(1)} d_t^{(1)} (1-\omega_{kt}^{(1)}) x_k v_t = \vartheta_s^{(1)}(x, v), \tag{2.15}$$

where $v_k = d_k^{(2|1)} y_k$ and $d_k^{(2|1)} = a_k^{(2|1)} / \pi_k^{(2|1)}$ is the conditional second-phase weight. It is seen from (2.15) that $c_s(x, y)$ is computed from the first-phase sampling variance estimator.

It remains to evaluate $\boldsymbol{z}_k$. We have $f(\boldsymbol{A}_b) = \sum b_k \pi_k^{(2|1)} y_k + \sum (b_k^{(1)} - b_k \pi_k^{(2|1)}) \hat{y}_k^*(\boldsymbol{b})$ with $\hat{y}_k^*(\boldsymbol{b}) = (1 - I_k) t_k + I_k \mu_k(\hat{\boldsymbol{\beta}}^{(N)}(\boldsymbol{b}))$ and $\hat{\boldsymbol{l}}_\beta^{(N)}(\boldsymbol{\beta}(\boldsymbol{b})) = \sum b_k \pi_k^{(2|1)} I_k \Phi_k(\boldsymbol{\beta}(\boldsymbol{b}))(y_k - \mu_k(\boldsymbol{\beta}(\boldsymbol{b}))) = 0$, where $\boldsymbol{b}$ is the second column of the matrix $\boldsymbol{A}_b$. Hence,

$$\boldsymbol{z}_k = \partial f(\boldsymbol{A}_b)/\partial \boldsymbol{b}_k \mid_{\boldsymbol{A}_b = \boldsymbol{A}_d} = \begin{cases} \hat{y}_k^* \\ g_k^{(N)}(y_k - \hat{y}_k^*), \end{cases} \tag{2.16}$$

with

$$g_k^{(N)} = \pi_k^{(2|1)} - \pi_k \hat{\boldsymbol{J}}_\theta^{(N)}(\boldsymbol{\beta}) [\hat{\boldsymbol{J}}_\beta^{(N)}(\hat{\boldsymbol{\beta}})]^{-1} \boldsymbol{\Phi}_k(\hat{\boldsymbol{\beta}}_a) I_k, \tag{2.17}$$

where $\hat{\boldsymbol{J}}_\theta^{(N)}(\boldsymbol{\beta}) = -\partial \hat{Y}^{(N)}(\boldsymbol{\beta})/\partial \boldsymbol{\beta}$, and $\hat{\boldsymbol{J}}_\beta^{(N)}(\boldsymbol{\beta}) = -\partial \hat{\boldsymbol{l}}_\beta^{(N)}(\boldsymbol{\beta})/\partial \boldsymbol{\beta}$.

Under the linear model

$$\mu_k = \boldsymbol{x}_k^T \boldsymbol{\beta} \text{ and } Var_m(y_k) = \sigma^2 / c_k, \tag{2.18}$$

for some specified constants c_k, $\Phi_k(\boldsymbol{\beta}) = \boldsymbol{x}_k c_k$, $\boldsymbol{\beta}$ is estimated by $\hat{\boldsymbol{\beta}}^{(N)} = \hat{\boldsymbol{Q}}_{(N)}^{-1} \sum a_k I_k c_k \boldsymbol{x}_k y_k$ with $\hat{\boldsymbol{Q}}_{(N)} = \sum a_k I_k c_k \boldsymbol{x}_k \boldsymbol{x}_k^T$. We have $\hat{\boldsymbol{J}}_\theta^{(N)}(\boldsymbol{\beta}) = \sum d_k^{(1)} (1 - a_k^{(2|1)}) I_k \boldsymbol{x}_k^T$, $\hat{\boldsymbol{J}}_\theta^{(N)}(\boldsymbol{\beta}) = \hat{\boldsymbol{Q}}_{(N)}$, and

$$g_k^{(N)} = \pi_k^{(2|1)} + \pi_k \sum_t d_t^{(1)}(1 - a_t^{(2|1)}) I_t \boldsymbol{x}_t^T \hat{\boldsymbol{Q}}_{(N)}^{-1} \boldsymbol{x}_k c_k I_k . \tag{2.19}$$

A variance estimator of the bias-adjusted estimator $\hat{Y}$ is given by (2.9) with $\boldsymbol{u}_k$ replaced by

$$\boldsymbol{z}_k = \partial f(\boldsymbol{A}_b)/\partial \boldsymbol{b}_k \,|_{\boldsymbol{A}_b = \boldsymbol{A}_d} = \begin{cases} \hat{y}_k^* \\ g_k (y_k - \hat{y}_k^*), \end{cases} \tag{2.20}$$

where

$$g_k = 1 - \hat{\boldsymbol{J}}_\theta(\boldsymbol{\beta})[\hat{\boldsymbol{J}}_\beta(\hat{\boldsymbol{\beta}})]^{-1} \boldsymbol{\Phi}_k(\hat{\boldsymbol{\beta}}) I_k ,$$

$f(\boldsymbol{A}_d) = \hat{Y}$, $\hat{\boldsymbol{J}}_\theta(\boldsymbol{\beta}) = -\partial \hat{Y}/\partial \boldsymbol{\beta}$, and $\hat{\boldsymbol{J}}_\beta(\boldsymbol{\beta}) = -\partial \hat{\boldsymbol{l}}_\beta(\boldsymbol{\beta})/\partial \boldsymbol{\beta}$. Under the linear model, we have $\hat{y}_{2k}^* = \boldsymbol{x}_k^T \hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{Q}}^{-1} \sum d_k I_k c_k \boldsymbol{x}_k y_k$ with $\hat{\boldsymbol{Q}} = \sum d_k I_k c_k \boldsymbol{x}_k \boldsymbol{x}_k^T$, and

$$g_k = 1 + \sum_t (d_t^{(1)} - d_t) I_t \boldsymbol{x}_t^T \hat{\boldsymbol{Q}}^{-1} \boldsymbol{x}_k c_k I_k . \tag{2.21}$$

2.3 Simulation Study

We conducted a small simulation study to examine the performances of the estimators $\hat{Y}^{(N)}$ and $\hat{Y}$ and associated variance estimators, when substitution imputation or ratio imputation is used. The more general case of values from administrative file when available and regression imputation otherwise is not studied in this section. We first generated a finite population $\{\boldsymbol{y}_1, ..., \boldsymbol{y}_k, ..., \boldsymbol{y}_N\}$, with $\boldsymbol{y}_k = (y_{1k}, y_{2k}, y_{3k})^T$, of size $N = 393$ from the following models: $y_{1k} = x_k + x_k^{1/2}\varepsilon_k$, $y_{2k} = 1.2x_k + x_k^{1/2}\varepsilon_k$, and $y_{3k} = 5 + 1.2x_k + x_k^{1/2}\varepsilon_k$, where ε_k are independent observations generated from $N(0,1)$, and the fixed x_k are the "number of beds" for the Hospitals population studied in Valliant *et al.* (2000, p.424-427). We stratified the population into two strata with 272 units k having $x_k \le 350$ in stratum 1 and 121 units k with $x_k > 350$ in stratum 2. We selected $R = 5{,}000$ stratified two-phase simple random samples of sizes $m_1 = m_2 = 50$ and $n_1 = n_2 = 15$. Our vector parameter of interest is the finite population total $\boldsymbol{\theta}_N = (Y_1, Y_2, Y_3)^T$. Values for the first-phase units not sampled in the second phase are mass imputed. For each variable l, $l = 1,2,3$, four estimators are considered. Two estimators used the substitution imputation method: the naïve estimator $\hat{\theta}_{lS}^{(N)} = \sum d_k^{(1)} \{a_k y_{lk} + (1 - a_k) x_k\}$, and the corresponding design-based estimator $\hat{\theta}_{lS} = \sum d_k y_{lk} + \sum (d_k^{(1)} - d_k) x_k$, $l = 1,2,3$. The other two estimators used the ratio imputation method: the naïve estimator $\hat{\theta}_{lR}^{(N)} = \sum d_k^{(1)} \{a_k y_{lk} + (1 - a_k) x_k \hat{\beta}_l^{(N)}\}$, and the associated design-based estimator $\hat{\theta}_{lR} = \sum d_k y_{lk} + \sum (d_k^{(1)} - d_k) x_k \hat{\beta}_l$, where $\hat{\beta}_l^{(N)} = \sum a_k y_{lk} / \sum a_k x_k$ and $\hat{\beta}_l = \sum d_k y_{lk} / \sum d_k x_k$. The estimator $\hat{\theta}_{lR}$ reduces to the two-phase ratio estimator: $\hat{Y}_R = \hat{Y}(\hat{X}^{(1)} / \hat{X})$.

Let $\hat{\theta}$ denote an estimator of a population total θ_N and $\vartheta(\hat{\theta})$ be the associated variance estimator. We calculated the simulated relative bias of $\hat{\theta}$ and $\vartheta(\hat{\theta})$ as $RB(\hat{\theta}) = (\hat{\theta} - \theta_N)/\theta_N$ and $RB\{\vartheta(\hat{\theta})\} = \{\vartheta(\hat{\theta}) - MSE(\hat{\theta})\}/MSE(\hat{\theta})$, where $\hat{\theta} = R^{-1} \sum_{r=1}^R \hat{\theta}_r$ is the mean of the estimates $\hat{\theta}_r$ from the simulated samples $r = 1, ..., R$ and $MSE(\hat{\theta}) = R^{-1} \sum_{r=1}^R (\hat{\theta}_r - \theta_N)^2$ is the simulated mean squared error (MSE). We calculated $RB(\hat{\theta})$, $RB\{\vartheta(\hat{\theta})\}$ and mean squared error (MSE) ratios for each component of the vector $(\hat{\boldsymbol{\theta}}_1^T, \hat{\boldsymbol{\theta}}_2^T, \hat{\boldsymbol{\theta}}_3^T)^T$ with $\hat{\boldsymbol{\theta}}_l = (\hat{\theta}_{lS}^{(N)}, \hat{\theta}_{lS}, \hat{\theta}_{lR}^{(N)}, \hat{\theta}_{lR})^T$ ($l = 1,2,3$), and those values are reported in Table 1.

It is clear from Table 1 that the substitution naïve estimator $\hat{\theta}_S^{(N)}$ can lead to large relative bias and loss in efficiency when the model generating the variable of interest is different from the model generating the imputed values. For example, $\hat{\theta}_S^{(N)}$ performs poorly for the variables 2 and 3. On the other hand, the design based estimator $\hat{\theta}$ performs well regardless of the population model and imputation method. The variance estimator ϑ_L for $\hat{\theta}_S^{(N)}$ leads to serious

underestimation of MSE for variables 2 and 3. On the other hand, variance estimator ϑ_L for $\hat{\theta}$ performs well with small RB. The naïve estimator $\hat{\theta}_R^{(N)}$ under ratio imputation is more robust to model deviations than $\hat{\theta}_S^{(N)}$.

3. Missing Responses

In this section, we consider the case of missing second-phase responses, y. Imputation is performed on all the first-phase missing responses. The resulting first-phase sample is complete. An estimator of the finite population total $Y = \sum y_k$ is given by

$$\hat{Y}(\tau) = \sum d_k o_k \tau_{\theta,k} y_k + \sum (d_k^{(1)} - d_k o_k \tau_{\theta,k}) \hat{y}_k^*, \tag{3.1}$$

for specified $\tau_{\theta,k}$, where $o_k = 1$ if y_k is observed, and $o_k = 0$ otherwise. Here, the imputed value $\hat{y}_{2k}^*$ is given by $\hat{y}_{2k}^* = \mu_k(\hat{\boldsymbol{\beta}})$ where the estimator $\hat{\boldsymbol{\beta}}$ is obtained as solution to the estimating equations

$$\hat{\boldsymbol{l}}_\beta(\boldsymbol{\beta}) = \sum d_k o_k \tau_{\beta,k} I_k \boldsymbol{\Phi}_k(\boldsymbol{\beta})(y_k - \mu_k(\boldsymbol{\beta})) = 0, \tag{3.2}$$

for specified $\tau_{\beta,k}$.

In the case of complete response, the choice of $\tau_{\theta,k} = \tau_{\beta,k} = 1$ gives the design-based estimator $\hat{Y} = \sum d_k y_k + \sum (d_k^{(1)} - d_k) \hat{y}_k^*$, and the choice of $\tau_{\theta,k} = \tau_{\beta,k} = \pi_k^{(2|1)}$ leads to the naïve estimator $\hat{Y}^{(N)} = \sum d_k^{(1)} [a_k^{(2|1)} y_k + (1 - a_k^{(2|1)}) \hat{y}_k^*]$, studied in section 2. In the case of incomplete response, the choice of $\tau_{\theta,k} = \tau_{\beta,k} = \xi_k^{-1}$ gives the design-response based estimator $\hat{Y}_\xi = \sum d_k (o_k / \xi_k) y_k + \sum (d_k^{(1)} - d_k (o_k / \xi_k)) \hat{y}_k^*$, where $\xi_k = E_r(o_k)$ and E_r denotes model expectation under response mechanism.

We suppose that the parameter of interest is $\theta_N = E_p E_r(\hat{Y}(\tau))$. Let $\boldsymbol{d}_k = (d_{1k}, d_{2k}, d_{3k})^T$, where $d_{1k} = d_k^{(1)}$, $d_{2k} = d_k$ and $d_{3k} = d_k o_k$. Then the Demnati–Rao (DR) variance estimator (Demnati and Rao, 2007) is simply given by

$$\vartheta_{DR}(\hat{Y}(\tau)) = \vartheta(z), \tag{3.3}$$

where $\vartheta(\boldsymbol{u})$, in operation notation, is the variance estimator of the linear combination $\hat{U} = \sum \boldsymbol{u}_k^T \boldsymbol{d}_k$, $\boldsymbol{u}_k = (u_{1k}, u_{2k}, u_{3k})^T$ a vector of constants, $\boldsymbol{z}_k = \partial f(\boldsymbol{A}_b) / \partial \boldsymbol{b}_k \,|_{\boldsymbol{A}_b = \boldsymbol{A}_d}$, $f(\boldsymbol{A}_d) = \hat{Y}(\tau)$, $\boldsymbol{A}_d$ is a $3 \times N$ matrix with k^{th} column $\boldsymbol{d}_k$, and $\boldsymbol{A}_b$ is a $3 \times N$ matrix of arbitrary real numbers with k^{th} column $\boldsymbol{b}_k = (b_{1k}, b_{2k}, b_{3k})^T$. We have,

$$z_k = \partial f(\boldsymbol{A}_b) / \partial \boldsymbol{b}_k \,|_{\boldsymbol{A}_b = \boldsymbol{A}_d} = \begin{cases} \hat{y}_k^* \\ 0 \\ g_k(\tau)(y_k - \hat{y}_k^*), \end{cases} \tag{3.4}$$

with

$$g_k(\tau) = \tau_{\theta,k} - \tau_{\beta,k} \hat{\boldsymbol{J}}_\theta(\hat{\boldsymbol{\beta}}) [\hat{\boldsymbol{J}}_\beta(\hat{\boldsymbol{\beta}})]^{-1} \boldsymbol{\Phi}_k(\hat{\boldsymbol{\beta}}) I_k \tag{3.5}$$

where $\hat{\boldsymbol{J}}_\theta(\boldsymbol{\beta}) = -\partial \hat{Y}(\tau) / \partial \boldsymbol{\beta}$, and $\hat{\boldsymbol{J}}_\beta(\boldsymbol{\beta}) = -\partial \hat{\boldsymbol{l}}_\beta(\boldsymbol{\beta}) / \partial \boldsymbol{\beta}$. Under linear model with $\tau_{\theta,k} = \tau_{\beta,k} = \tau_k$, we have $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{Q}}^{-1} \sum d_k \tau_k o_k I_k c_k \boldsymbol{x}_k y_k$, $\hat{\boldsymbol{Q}} = \sum d_k \tau_k o_k I_k c_k \boldsymbol{x}_k \boldsymbol{x}_k^T$, and

$$g_k(\tau) = \tau_k \{1 + \sum (d_t^{(1)} - d_t o_t \tau_t) I_t \boldsymbol{x}_t^T \hat{\boldsymbol{Q}}^{-1} \boldsymbol{x}_k c_k I_k\}. \tag{3.6}$$

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

$$\vartheta(\boldsymbol{u}) = \vartheta_s(\boldsymbol{u}_s) + \vartheta_r(u_o), \tag{3.7}$$

where $u_{k;o} = u_{3k}$ and $\boldsymbol{u}_{k;s} = (u_{1k}, u_{2k} + o_k u_{k;o})^T$.

The design based variance estimator $\vartheta_s(\boldsymbol{u}_s)$ is given by (2.9) with $\boldsymbol{u}_k$ replaced by $\boldsymbol{u}_{k;s}$.

The response estimated variance $\vartheta_r(u)$, assuming that the order of expectation can be interchange so that $E_p E_r = E_r E_p$, is given by

$$\vartheta_r(u) = \sum\sum d_k d_t \omega_{kt} o_k o_t (\hat{\xi}_{kt} - \hat{\xi}_k \hat{\xi}_t) \hat{\xi}_{kt}^{-1} u_k u_t, \tag{3.8}$$

where $\hat{\xi}_k = \hat{E}_r(o_k)$, $\hat{\xi}_{kt} = \hat{E}_r(o_k o_t)$, and $\hat{E}_r$ denotes the estimator of model expectation under a model on o_k. A variance estimator of $\hat{Y}(\tau)$ is then given by (3.7) with $\boldsymbol{u}_k$ replaced by $\boldsymbol{z}_k$ given by (3.4). Note that $\theta_N = Y$ if $\hat{Y}(\tau)$ is unbiased for Y under the assumed response mechanism and the sampling design.

4. Determination of "Optimal" Sample Sizes

4.1 Complete Response

Let y_k be the value obtained using a questionnaire, and t_k or x_k be the value obtained from administrative sources. We assume that obtaining values from administrative sources is much cheaper than obtaining values through questionnaire. We consider the case where a stratified first-phase sample of size $\boldsymbol{m} = (m_1, ..., m_h, ..., m_H)$ of t_k or x_k is obtained, and then from each first-phase sample $s_h^{(1)}$, a sub-sample s_h of size n_h is interviewed and the response y_k is recorded. Our interest is to find the "optimal" sample sizes $\boldsymbol{m} = (m_1, ..., m_h, ..., m_H)$ and $\boldsymbol{n} = (n_1, ..., n_h, ..., n_H)$ for estimating the finite population parameter θ_N that minimize the cost subject to constraints on the variances and the sample sizes.

Consider first the general form of an estimator in the complete response case:

$$\hat{U} = \sum d_k^{(1)} u_{1k} + \sum d_k u_{2k} = \sum \boldsymbol{u}_k^T \boldsymbol{d}_k \tag{4.1}$$

where $\boldsymbol{u}_k = (u_{1k}, u_{2k})^T$ and $\boldsymbol{d}_k = (d_k^{(1)}, d_k)^T$. If the conditional second-phase inclusion probabilities are constant, then

$$Cov_p(\boldsymbol{d}_k, \boldsymbol{d}_t) = \begin{pmatrix} (1-\omega_{kt}^{(1)})/\omega_{kt}^{(1)} & (1-\omega_{kt}^{(1)})/\omega_{kt}^{(1)} \\ (1-\omega_{kt}^{(1)})/\omega_{kt}^{(1)} & (1-\omega_{kt})/\omega_{kt} \end{pmatrix}. \tag{4.2}$$

Even when the conditional second-phase inclusion probabilities are random the right side of equation (4.2) is often used as an approximation to $Cov_p(\boldsymbol{d}_k, \boldsymbol{d}_t)$. Using (4.2), the sampling variance of the estimated total $\hat{U}$ is given by

$$Var_p(\boldsymbol{u}) = \sum\sum \boldsymbol{u}_k^T Cov_p(\boldsymbol{d}_k, \boldsymbol{d}_t) \boldsymbol{u}_t. \tag{4.3}$$

Substituting (4.2) into (4.3), we get

$$\begin{aligned} Var_p(\boldsymbol{u}) = & \sum\sum[(1-\omega_{kt}^{(1)})/\omega_{kt}^{(1)}] u_{1k} u_{1t} \\ & + \sum\sum[(1-\omega_{kt})/\omega_{kt}] u_{2k} u_{2t} \\ & + 2\sum\sum[(1-\omega_{kt}^{(1)})/\omega_{kt}^{(1)}] u_{1k} u_{2t}. \end{aligned} \tag{4.4}$$

Under stratified simple random sampling (STSRS) at both stages: $s_h^{(1)}$ is a simple random sample of size m_h from stratum h and s_h is simple random sample of size n_h from $s_h^{(1)}$, we have $\pi_{hk}^{(1)} = m_h / N_h$, $\pi_{hk} = n_h / N_h$, and for $k \neq t$, $\pi_{hkt}^{(1)} = m_h(m_h - 1)/[N_h(N_h - 1)]$, and $\pi_{hkt} = n_h(n_h - 1)/[N_h(N_h - 1)]$. We get from (4.4),

$$Var_p(\boldsymbol{u}) = \sum_h N_h^2 m_h^{-1}(1 - m_h/N_h)[S_{hu_1u_1} + 2S_{hu_1u_2}] + \sum_h N_h^2 n_h^{-1}(1 - n_h/N_h) S_{hu_2u_2} \tag{4.5}$$

where $S_{hxy} = \sum J_{hk}(x_k - \overline{X}_h)(y_k - \overline{Y}_h)/(N_h - 1)$, $\overline{X}_h = \sum J_{hk} x_k / N_h$ and J_{hk} is the stratum membership indicator variable for element k.

4.2 Optimal Sample Sizes: STSRS

Consider the complete response case with p characteristics of interest $y_1,...,y_p$, and assume that the estimator $\hat{U}$ is unbiased under the sampling design. Then, under STSRS, it follows from (4.5) that we can express $Var_p(\hat{U}_j)$, for the j^{th} variable y_j, as

$$Var_p(\hat{U}_j) = v_{j0} + \sum_h (v_{jh1}/m_h + v_{jh2}/n_h), \ j = 1,..., p \tag{4.6}$$

where $v_{j0} = -\sum_h N_h S_{hjuu}$, $v_{hj1} = N_h^2 (S_{hju_1u_1} + 2S_{hju_1u_2})$, $v_{hj2} = N_h^2 S_{hju_2u_2}$, $u_{jk} = u_{j1k} + u_{j2k}$. We determine the optimal m_h and n_h by mathematical programming methods such that the cost

$$C = c_0 + \sum_h (c_{h1} m_h + c_{h2} n_h) \tag{4.7}$$

is minimized subject to constraint on the variances

$$Var_p(\hat{U}_j) \le V_j, \ j = 1,..., p, \tag{4.8}$$

and constraint on samples sizes and population sizes

$$2 \le n_h \le m_h \le N_h, \ h = 1,..., H, \tag{4.9}$$

where c_0 is the overhead cost, c_{h1} and c_{h2} are the cost per element in stratum h for the first-phase and second-phase sample respectively, and V_j is the specified tolerance. For example, one could specify upper limit, ϕ_j, on the coefficient of variation of $\hat{\theta}_j$ so that $V_j = (\phi_j E_p \hat{\theta}_j)^2$. We assume that $c_{h1} << c_{h2}$.

We consider the bias-adjusted estimator $\hat{Y}_j = \sum d_k y_{jk} + \sum (d_k^{(1)} - d_k)\hat{y}_{jk}^*$ of the total Y_j, given by (2.6) with $\hat{\boldsymbol{\beta}}$ obtained as solution to the estimating equations $\hat{\boldsymbol{l}}_{j\beta}(\boldsymbol{\beta}) = \sum d_k I_{jk} \boldsymbol{\Phi}_{jk}(\boldsymbol{\beta})(y_{jk} - \mu_{jk}(\boldsymbol{\beta})) = 0$. In this case

$$\boldsymbol{u}_{jk} = \begin{cases} y_{jk}^* \\ y_{jk} - y_{jk}^* \end{cases} \equiv \begin{cases} u_{j1k} \\ u_{j2k} \end{cases} \tag{4.10}$$

with $y_{jk}^* = (1 - I_{jk}) t_{jk} + I_{jk} \mu_{jk}(\boldsymbol{\beta}_{jN})$, and $\boldsymbol{\beta}_{jN}$ is the solution to $\boldsymbol{l}_{j\beta}(\boldsymbol{\beta}) = \sum I_{jk} \boldsymbol{\Phi}_{jk}(\boldsymbol{\beta})(y_{jk} - \mu_{jk}(\boldsymbol{\beta})) = 0$.

4.3 Empirical Study

We considered the case of a single characteristic, y, and generated a finite population $\{y_1,...,y_N\}$ of size $N = 393$ from the model

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

with independent errors ε_k generated from $N(0,1)$, $\mu_k = E_m(y_k) = \alpha + \beta x_k$ and specified constants (α, β, σ), where the fixed x_k are the "number of beds" in hospital k for the hospitals population. We stratified the population into two strata with 272 units k having $x_k \le 350$ in stratum 1 and 121 units k with $x_k > 350$ in stratum 2. For the cost, we set $c_0 = 0$, $c_{h2} = c_2 = 1$, and two different costs for $c_{h1} = c_1$: 0.1 and 0.5. We set $\phi = 0.05$ for the tolerances. Two estimators are considered $\hat{Y}_S = \sum d_k y_k + \sum (d_k^{(1)} - d_k) x_k$ and $\hat{Y}_R = \sum d_k y_k + \sum (d_k^{(1)} - d_k) x_k \hat{\beta}$ with $\hat{\beta} = \sum d_k y_k / \sum d_k x_k$. We have $\boldsymbol{u}_k = (x_k, y_k - x_k)^T$ for $\hat{Y}_S$ and $\boldsymbol{u}_k = (x_k \beta_N, y_k - x_k \beta_N)^T$ for $\hat{Y}_R$, where $\beta_N = \sum y_k / \sum x_k$.

We repeated the optimization process for different value of $\sigma,(0,..,15)$ keeping $(\alpha,\beta)=(0,1)$ and (0,1.2). Figures 1 and 2 provide the minimum cost for $\beta=1.0$ under the two-phase design for $c_1=0.5$ and $c_1=0.1$ respectively. The results under substitution and ratio imputation are similar.

For comparison, the minimum cost under stratified single phase sampling, using the unbiased Horvitz-Thompson (HT) estimator are also presented. It is seen from Figure 1 that when the cost ratio c_2/c_1 is small ($c_2/c_1=2$), minimum cost under two-phase is smaller than under one-phase sampling only for small values of σ (≤ 2). On the other hand, Figure 2 shows that when the cost ratio is large ($c_2/c_1=10$), minimum cost is smaller under two-phase sampling for a much wider ranger of σ (≤ 10). Figures 3 and 4 report the minimum cost for the case of $\beta=1.2$ under substitution and ratio imputation for $c_1=0.5$ and $c_1=0.1$ respectively.

4.4 Missing Responses

In the case of missing responses, y, in the second-phase sample, the variance of the estimator of general form $\hat{U}=\sum d_k^{(1)}u_{1k}+\sum d_k o_k u_{2k}$ is given by

$$Var(\hat{U})=V_s(\boldsymbol{u})+V_r(u_2),$$

with

$$\begin{aligned}V_s(\boldsymbol{u})&=\sum\sum[(1-\omega_{kt}^{(1)})/\omega_{kt}^{(1)}]u_{1k}u_{1t}\\&+\sum\sum[(1-\omega_{kt})/\omega_{kt}]\xi_{kt}u_{2k}u_{2t}\\&+2\sum\sum[(1-\omega_{kt}^{(1)})/\omega_{kt}^{(1)}]\xi_t u_{1k}u_{2t},\end{aligned}\tag{4.11}$$

and

$$V_r(u_2)=\sum\sum(\xi_{kt}-\xi_k\xi_t)u_{2k}u_{2t}.\tag{4.12}$$

We assume independent response mechanism. Under stratified simple random sampling at both stages, we have

$$\begin{aligned}V_s(\boldsymbol{u})&=\textstyle\sum_h N_h^2/m_h(1-m_h/N_h)[S_{hu_1u_1}+2S_{hu_1\ddot{u}_2}]\\&+\textstyle\sum_h N_h^2/n_h(1-n_h/N_h)[S_{h\ddot{u}_2\ddot{u}_2}+\sum J_{hk}\xi_k(1-\xi_k)u_{h1k}u_{h2k}]\end{aligned}\tag{4.13}$$

where $\ddot{x}_k=x_k\xi_k$, $u_k=u_{1k}+u_{2k}$, and $\vec{u}_k=u_{1k}+\ddot{u}_{2k}$. It follows from (4.11), (4.12) and (4.13), that we can express $Var(\hat{U})$ as (4.6) with,

$$v_{hj1}=N_h^2(S_{hu_1u_1}+2S_{hu_1\ddot{u}_2}),$$

$$v_{hj2}=N_h^2[S_{h\ddot{u}_2\ddot{u}_2}+\sum J_{hk}u_{h1k}u_{h2k}\xi_k(1-\xi_k)],$$

and $$v_{j0}=-\textstyle\sum_h N_h[S_{hu\vec{u}}+\sum J_{hk}u_{h1k}u_{h2k}\xi_k(1-\xi_k)]+\sum\xi_k(1-\xi_k)u_{2k}^2.$$

References

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 (with comments)". *Third International Conference on Establishment Surveys*, Montréal, Canada, 916-925.

Demnati, A. and Rao, J.N.K. (2008), "Linearized Variance Estimation from Simulated Census Data". *Proceedings of the Survey Methods Section,* Statistical Society of Canada.

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

Table 1: Simulation Results

Imputation Method	Estimator	Variable	$RB(\hat{\theta})$	$MSE(\hat{Y}^{(N)})/MSE(\hat{Y})$	$RB\{\vartheta(\hat{\theta})\}$
Substitution	Naïve	1	.0001	.94	-.01
		2	-.1181	9.85	-.93
		3	-.1228	10.66	-.93
	Design-based	1	.00004	1	-.01
		2	.0003	1	-.001
		3	.0005	1	-.001
Ratio	Naïve	1	-.0004	.99	-.010
		2	.0004	.99	-.011
		3	-.0023	.99	-.006
	Design-based	1	.00004	1	-.011
		2	.0003	1	-.010
		3	.0006	1	-.002

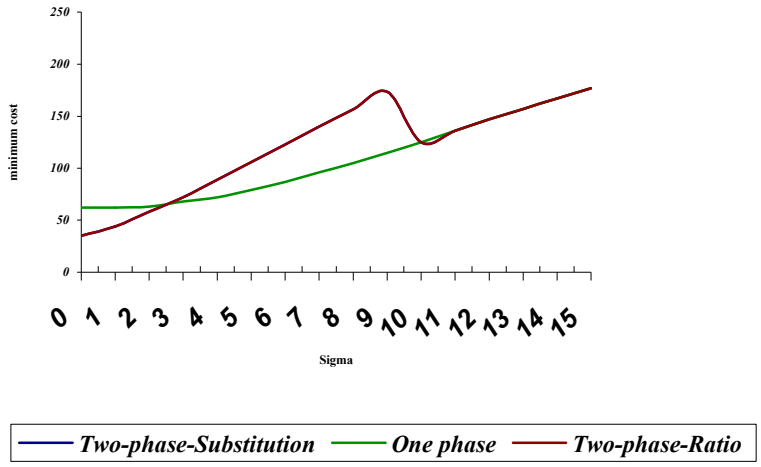

Figure 1: minimum cost for $(\beta = 1.0, c_1 = 0.5)$

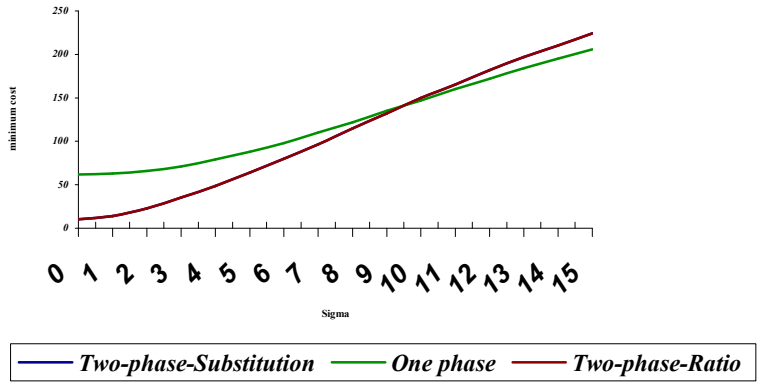

Figure 2: minimum cost for $(\beta = 1.0, c_1 = 0.1)$

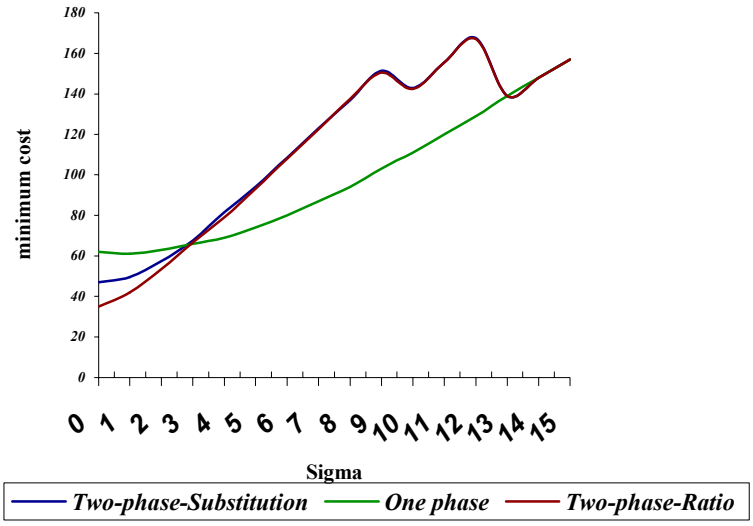

Figure 3: minimum cost for $(\beta = 1.2, c_1 = 0.5)$

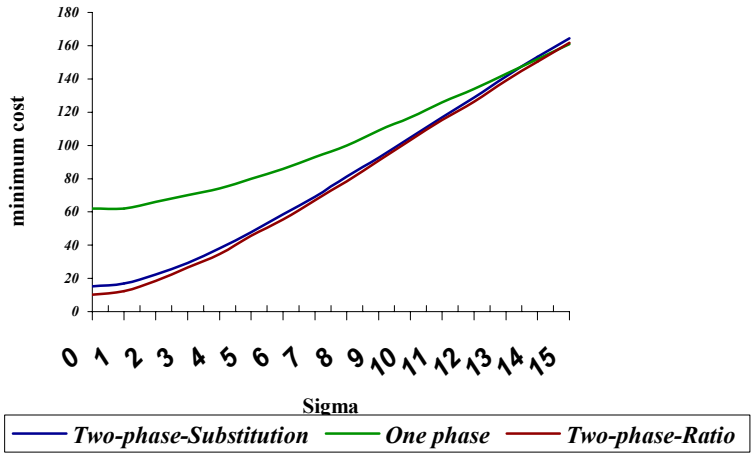

Figure 4: minimum cost for $(\beta = 1.2, c_1 = 0.1)$