Dealing with Nonresponse using follow-up

Hidiroglou, M.A. and V. Estevao
Statistical Research and Innovation Division, Statistics Canada,
16th Floor, R. H. Coats Bldg., Tunney's Pasture, Ottawa, Ontario, K1A0T6

Abstract

Non-response in surveys is usually addressed using imputation or weighting adjustments, given that no follow-up sample has been selected. These procedures are reasonable if the respondents and non-respondents are similar. If they are not, these procedures may lead to serious biases. A follow-up sample of the non-respondents will eliminate (attenuate) any such bias. However, since not all units selected in the non-response sample will respond, this leads to a three-phase sampling design. In this paper we provide weighting procedures that properly account for this, as well as the associated population and estimated variances that reflect the weighting. Also, given that a sample of the non-respondents will be followed up, we provide the allocation between the initial sample and follow-up sample size given cost constraints.

Key words: Nonresponse; follow-up; three-phase sampling; allocation

1. Introduction

The quality of surveys is affected by several types of errors. One of the most important ones is nonresponse. Nonresponse may introduce biases into the estimates that the surveyor may not even be aware of. Response rates may us an idea of the extent of the bias (upper bound): generally speaking the higher the response rates the more likely that the bias is not important. However, one can never be sure as it is easy to construct scenarios where the bias is actually smaller for lower response rates.

There are two approaches for attenuating the potential bias associated with nonresponse. The first one is applied at the design stage by ensuring that a subsample of the nonrespondents is followed up: Hansen and Hurwitz (1946) published the first paper on this procedure. The second approach is to incorporate auxiliary information that is related to the variable of interest in the estimation of the parameters. This auxiliary information may be in the form of estimating the probability of response for each unit included in the sample as in Fuller et al. (1994), or as direct auxiliary data incorporated in the estimation as in Lundström and Särndal. (1999), or as combination of both as in Kott and Chang (2010). It should be noted that the dual use of a follow-up sample and auxiliary data is not addressed in the literature.

In this paper, we recognize that nonresponse will also occur in a follow-up. Given expected response rates for the respondents and the followed up sample of nonrespondents, we develop in section 2 an optimum allocation scheme that addresses the allocation of the sample ignoring auxiliary data. As suggested by Little (1986) we split the nonresponding sample into response homogeneity groups (based on past knowledge) so as minimize as much as possible nonresponse bias in the follow-up sample. We apply a simplified version of the procedure to the data set of Hansen and Hurwitz (1946). Given that we may not always be able to split the sample into response homogeneity groups, we study the impact of also using auxiliary data to reduce the bias in section 4. We summarize our findings in section 5.

2. Sampling and Estimation with a follow-up sample

A sample *s* of size *n* is selected by simple random sampling without replacement from a population U of size *N*. Suppose that $n_1 < n$ units respond to the survey, while the remaining ones do not. Hansen and Hurwitz (1946) were the first to suggest sub-sampling the nonrespondents. They assumed that all followed-up units would respond. This assumption ensures that the resulting estimate will be unbiased. However, as this is often not the case, the resulting estimate may be biased.

One way to attenuate this bias is to split the set of nonresponding units in the sample into response homogeneity groups (strata) that reflect similar propensity for response. Suppose that the nonresponding set is split into such $L-1$ strata. We can then view the sample *s* as having been stratified into *L* strata with each stratum being of size N_h. That is $s=\bigcup_{h=1}^{L} s_h$ with n_h units sampled in stratum *h* ($h=1,...,L$): note that $n=\sum_{h=1}^{L} n_h$. The first stratum U_1 corresponds to the subset of the sampled units that have responded, and the remaining $L-1$ strata U_h , $(h=2,\ldots,L)$ represent homogeneous groups of nonrespondents. The portion of the sample *s* that responds is denoted as s_1, while $s_2=\bigcup_{h=2}^{L} s_{2h}$ corresponds to the set that does not respond.

Stratified second-phase samples s_{3h}, of m_h units, $h=2,\ldots,L$ are selected with SRSWOR from the n_h nonrespondents in each of the $(L-1)$ nonresponding samples s_{2h}. Assume that b_h units, $0<b_h \leq m_h$, respond to the follow-up. We denote the responding subsets of s_{3h} as s_{4h}. The sample selection process is shown below in Figure 1.

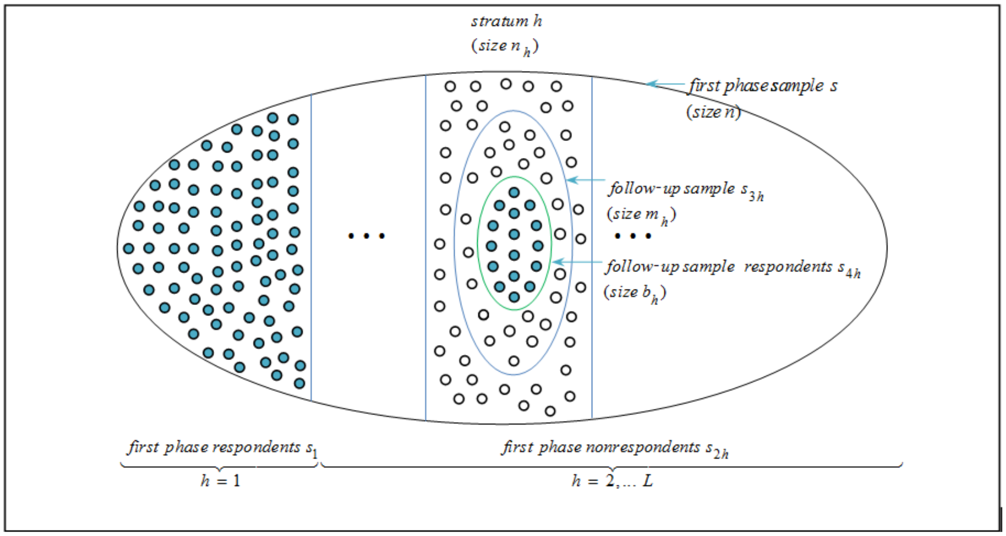
stratum h
(size n_h)
first phase sample s
(size n)
follow-up sample s_{3h}
(size m_h)
follow-up sample respondents s_{4h}
(size b_h)
first phase respondents s_1
$h = 1$
first phase nonrespondents s_{2h}
$h = 2,\ldots L$

Figure 1: The respondents and nonrespondents are shown for h = 1 and the general stratum *h*. The respondents are represented by blue dots, while the nonrespondents are shown as circles (with no colour).

Proceeding as in Rao (1973), we assume that:

- $n = \nu\, N$ where ν is the first-phase sampling fraction such that $0 < \nu \le 1$,
- $m_h = \nu_h\, n_h$, where ν_h are the second-phase sampling fractions, with $\nu_h = 1$ for $h = 1$ and $0 < \nu_h \le 1$ for $h = 2, \dots, L$.

The parameter of interest is the population total $Y = \sum_{k \in U} y_k$. An estimator of *Y* is

$$\begin{aligned} \hat{Y}_3 &= \frac{N}{n}\left(\sum_{k \in s_1} y_k + \sum_{h=2}^{L} \frac{n_h}{m_h} \frac{m_h}{b_h} \sum_{k \in s_{4h}} y_k \right) \\ &= N\left(w_1\, \bar{y}_1 + \sum_{h=2}^{L} w_h\, \bar{y}_{4h} \right) \end{aligned} \qquad (1)$$

where $w_h = n_h / n$, $\bar{y}_1$ and $\bar{y}_{4h}$ $(h = 2, \dots, L)$ are the means of the responding sets.

If we assume that the non-responding units in s_{3h} are missing completely at random (MCAR), then s_{4h} is an SRS sample of size b_h from U_h $(h = 2, \dots, L)$, given n_h and b_h. It can be shown that $\hat{Y}_3$ is an unbiased estimator of *Y* using the conditional expectation $E\left(\hat{Y}_3\right) = E_1\, E_2\left(\hat{Y}_3\right)$. Noting that $E_2\left(\bar{y}_{4h}\right) = \bar{Y}_h$, we have that

$$E_2\left(\hat{Y}_3\right) = \frac{N}{n}\left(n_1\, \bar{Y}_1 + \sum_{h=2}^{L} n_h\, \bar{Y}_h \right) = N\left(w_1\, \bar{Y}_1 + \sum_{h=2}^{L} w_h\, \bar{Y}_h \right) \qquad (2)$$

where $\bar{Y}_h$ are the population means of the sets U_h. Hence, $E\left(\hat{Y}_3\right) = N\left(W_1\, \bar{Y}_1 + \sum_{h=2}^{L} W_h\, \bar{Y}_h \right) = N\bar{Y} = Y$ noting that $E_1\left(w_h\right) = W_h$ where $W_h = N_h / N$.

The population variance of $\hat{Y}_3$ is obtained using

$$V\left(\hat{Y}_3\right) = V_1\, E_2\left(\hat{Y}_3\right) + E_1\, V_2\left(\hat{Y}_3\right) \qquad (3)$$

The first term of (3) is given by $V_1\, E_2\left(\hat{Y}_3\right) = N^2\left(\sum_{h=1}^{L} w_h \bar{Y}_h \right)$. The second term of (3) is derived as follows. Noting that

$$V_2\left(\hat{Y}_3\right) = N^2\left[w_1^2\left(\frac{1}{n_1} - \frac{1}{N_1} \right) S_1^2 + \sum_{h=2}^{L} w_h^2\left(\frac{1}{n_h} - \frac{1}{N_h} \right) S_h^2 \right]$$

we have

$$E_1 V_2\left(\hat{Y}_3\right) \approx N^2\left[\left\{E_1\left(\frac{w_1}{n_1}\right)-E_1\left(\frac{w_1^2}{N_1}\right)\right\}S_1^2+\sum_{h=2}^{L}\left\{E_1\left(\frac{w_h^2}{E_3\left(b_h\right)}\right)-E_1\left(\frac{w_h^2}{N_h}\right)\right\}S_h^2\right]$$

where $E_3\left(b_h\right)=m_h\, r_h^*=n_h\, \nu_h\, r_h^*$ is the expectation over b_h given n_h(or m_h), and r_h^* are the *anticipated response rates* within stratum *h*. The resulting variance $V\left(\hat{Y}_3\right)$ is given by

$$V\left(\hat{Y}_3\right)=N\left\{\left(\frac{1}{\nu}-1\right)S^2+\sum_{h=2}^{L}\frac{1}{\nu}\left(\frac{1}{\nu_h\, r_h^*}-1\right)W_h\, S_h^2\right\} \tag{4}$$

where $S^2=\sum_{k\in U}\left(y_k-\overline{y}_U\right)^2/(N-1)$ with $\overline{y}_U=\sum_{k\in U}y_k\,/\,N$, $S_h^2=\sum_{k\in U_h}\left(y_k-\overline{y}_{U_h}\right)^2/\left(N_h-1\right)$ with $\overline{y}_{U_h}=\sum_{k\in U_h}y_k\,/\,N_h$ and $W_h=N_h\,/\,N$ for $h=1,\ldots,L$.

Note that if complete response occurs in the follow-up sample ($r_h^*=1$ for $h=2,\ldots,L$), then formula (4) reduces to formula (2) given in Rao (1973).

There are costs associated with contacting and processing the initial sample and the follow-up sample. Let c_0 be the contact cost for each of the initial n units and c_1 be the unit cost of processing the n_1 respondents from this sample. Similarly, for the follow-up sample in strata $h=2,\ldots,L$, we let c_{2h} represent the unit contact cost of the m_h units and c_{3h} the cost of processing each of the b_h respondents. This means that the overall cost of taking the sample is

$$C\;=\;n\,c_0+n_1\,c_1+\sum_{h=2}^{L}(\,m_h\,c_{2h}+\,b_h\,c_{3h}) \tag{5}$$

However, this cost expression is a random quantity because n_1, $m_h=n_h\,\nu_h$ and b_h ($h=2,\ldots,L$) are random variables. We therefore need to work with the expected cost $C^*=E(C)$, where

$$C^*\;=\;N\,\nu\left(\,c_0+W_1\,c_1+\sum_{h=2}^{L}\,W_h\,\nu_h\,(\,c_{2h}+\,r_h^*\,c_{3h})\right) \tag{6}$$

This has the same form as the expected cost formula $n' c' + n' \sum W_h \, \nu_h \, c_h$ in Rao (1973). In this correspondence, it is clear that $N \nu = n'$ and $c_0 + W_1 c_1 = c'$. Our term $(c_{2h} + r_h^* c_{3h})$ is similar to the c_h term in Rao's expected cost formula but our sum only extends from $h = 2$ to $h = L$.

We can state the allocation problem as follows.

$$\text{Minimize} \quad N \left\{ \left(\frac{1}{\nu} - 1 \right) S^2 + \sum_{h=2}^{L} \frac{1}{\nu} \left(\frac{1}{\nu_h \, r_h^*} - 1 \right) W_h \, S_h^2 \right\}$$

with respect to ν, and $\nu_h \ \ (h = 2, \dots L)$

$$\text{subject to} \quad N \nu \left(c_0 + W_1 c_1 + \sum_{h=2}^{L} W_h \, \nu_h \, (c_{2h} + r_h^* \, c_{3h}) \right) = C^* \tag{7}$$

$$\text{and} \qquad 0 < \nu \le 1, \ \ 0 < \nu_h \le 1$$

To solve this problem we need to know or estimate the proportions W_h, the response rates r_h^* and the population variances S^2 and S_h^2. If N_h is not known, we estimate the proportions $W_h = N_h / N$ based on administrative data. Similarly, the anticipated response rates r_h^* can be estimated from previous surveys. Jeyaratnam et al. (1984) provide a methodology, due to Frayer (1979) to estimate S^2 and S_h^2. This is a multiphase analogue to the multistage variance estimation methodology in Rao (1975). The following conditions are required to apply for the procedure:

1. The first-phase sample is SRSWOR,
2. There exists an estimator, *y*, based on the second and subsequent phases of sampling which given the units selected in the first-phase, α, is an unbiased estimate of the total of these units, i.e., $y_1 + \ldots + y_n$,
3. As in (2) there exists an estimator, *z,* which is an unbiased estimator of $y_1^2 + \ldots + y_n^2$,
4. There exists an unbiased estimator of the variance of *y* given α, i.e., $\hat{V}\left(y | \alpha \right)$ is unbiased for $V\left(y | \alpha \right)$.

Conditions (1) to (4) are sufficient to establish that $\hat{S}^2 = \frac{1}{n-1} \left[z - \frac{y^2}{n} + \frac{1}{n} \hat{V}\left(y | \alpha \right) \right]$ is unbiased for S^2. In the context of a three-phase sample, this means that

$$\hat{S}^2 = \frac{1}{(n-1)} \left[\sum_{h=1}^{L} \frac{n_h}{m_h} z_h - \frac{1}{n} \left(\sum_{h=1}^{L} \frac{n_h}{m_h} y_h \right)^2 + \frac{1}{n} \sum_{h=1}^{L} \hat{V} \left(\frac{n_h}{m_h} y_h \mid \alpha \right) \right] \tag{8}$$

$$\hat{S}_h^2 = \frac{1}{(m_h - 1)} \left[z_h - \frac{y_h^2}{m_h} + \frac{1}{m_h} m_h^2 \left(\frac{1}{b_h} - \frac{1}{m_h} \right) \hat{S}_{3h}^2 \right] \qquad (9)$$

where

$$y_h = \frac{m_h}{b_h} \sum_{k \in s_{4h}} y_k \quad \text{and} \quad z_h = \frac{m_h}{b_h} \sum_{k \in s_{4h}} y_k^2$$

$$\hat{V}\left(\frac{n_h}{m_h} y_h \mid \alpha \right) = \frac{n_h (n_h - m_h)}{m_h (m_h - 1)} \left[z_h - \frac{y_h^2}{m_h} \right] + \frac{n_h (n_h - 1)}{m_h (m_h - 1)} m_h^2 \left(\frac{1}{b_h} - \frac{1}{m_h} \right) \hat{S}_{3h}^2$$

$$\hat{S}_{3h}^2 = \frac{1}{(b_h - 1)} \sum_{k \in s_{4h}} (y_k - \bar{y}_h)^2 \ \text{ with } \bar{y}_h = \frac{1}{b_h} \sum_{k \in s_{4h}} y_k \quad .$$

The allocation problem defined by (7) is a nonlinear programming problem since the function to minimize and the expected cost constraint are both nonlinear functions of the unknown sampling fractions. We can attempt a closed form solution by solving the problem without the bounds on the sampling fractions. This can be obtained by the method of Lagrangian multipliers or by applying the Cauchy-Schwarz inequality on the product of the variance and the cost constraint. Either method gives the following result.

$$v_h = S_h \sqrt{ \frac{ \left(c_0 + W_1 c_1 \right) }{ r_h^* \left(c_{2h} + r_h^* c_{3h} \right) \left(S^2 - \sum_{h=2}^{L} W_h S_h^2 \right) } } \qquad (10)$$

Equation (10) shows that optimal v_h does not depend on C^*. This solution is similar to result (8) in Rao (1973) if we take $r_h^* = 1$ and we match $c_{2h} + r_h^* c_{3h}$ to the c_h in his paper. The optimal value of v can be obtained as a function of the optimal v_h but we do not present it here since it is more complicated to write down.

.

If this solution for v and v_h satisfies $0 < v \le 1$ and $0 < v_h \le 1$ then it is the optimal solution to (7). Otherwise, we find the optimal solution with the bound constraints through non-linear programming methods such as the Trust region method.

3. Example

We now apply the results in section 2 to the example of a survey of 40,000 retail stores given in Hansen and Hurwitz (1946). The initial contact was via mail questionnaire and the follow-up was via face to face interviews. They assumed 100% response in the face to face interviews. The initial sample for mail out and the subsequent sample for follow-up of the nonrespondents were both obtained using simple random sampling without replacement.

Table 1 summarizes the information on the unit costs for the initial contact and the follow-up. Using this information, Hansen and Hurwitz determined the optimal sizes for the initial and follow-up samples to minimize total cost while achieving the same reliability as a sample design of 1000 units where all units respond without follow-up. Although this problem is the reverse of (8), the resulting solutions are identical. This is because minimizing cost subject to a fixed variance results in the same solution as minimizing variance subject to fixed costs.

Table 1: Data on the unit costs for computing sample sizes in Hansen and Hurwitz (1946).

Mail out cost per questionnaire	10 cents
Follow-up cost with filed follow-up	\$4.10
Processing costs for either contact method	40 cents

Their setup used $N = 40,000$, $h = 2$, $W_2 = 0.5$ and $r_2^* = 1$. Their problem was stated in terms of the unknown sample sizes. It can be expressed in the following equivalent form for comparison with our results which are based on the sampling fractions.

$$\begin{aligned}&\text{Minimize} \quad (N c_0 + N c_1 W_1)\nu + N c_2 W_2 \nu \nu_2 \\ &\text{with respect to } \nu, \text{ and } \nu_2 \\ &\text{Subject to} \quad \left(\frac{1}{\nu} - 1\right) S^2 + W_2 S_2^2 \frac{1}{\nu} (\frac{1}{\nu_2} - 1) = V\end{aligned} \qquad (11)$$

where $V = S^2/n$ is the estimated variance assuming complete response (no follow-up) for a sample with $n = 1000$. Their cost values were $c_0 = 0.1$, $c_1 = 0.4$ and $c_2 = 4.5$. Assuming $S_2^2 = S^2$, they obtained the following optimal solution: $\nu_2 = 0.36515$ and $\nu = 0.04675$.

These optimal sampling fractions result in $n = N\nu = 1870$ and $E(n_2) = n W_2 \nu_2 = 341$ with a total minimum cost of \$2095. We now look at the reverse problem (8) with $C^* = 2095$ and the same cost values and assumptions.

$$\begin{aligned} &\text{Minimize} \quad \left(\frac{1}{\nu}-1\right)+W_2\frac{1}{\nu}(\frac{1}{\nu_2}-1) \\ &\text{with respect to } \nu, \text{ and } \nu_2 \qquad (12) \\ &\text{Subject to} \quad (N c_0 + N c_1 W_1)\nu + N c_2 W_2 \nu \, \nu_2 = C^* \end{aligned}$$

We removed the constants N and $\hat{S}_2^2 = \hat{S}^2$ in the variance formula of (8) because they are present in each term. For this problem specification, we obtain the same optimal sampling fractions shown in (10), showing that minimizing variance subject to fixed cost gives the same solution as minimizing cost subject to fixed variance. Note that the optimal fractions satisfy the bound constraints. We can remove the bound constraints when they do not affect the solution to the problem. We can expand our analysis by removing the assumption $r_2^* = 1$ so we can compare the optimal allocations for different values of r_2^* and C^*. In the following problem definition, we have also chosen to remove the bound constraints.

$$\begin{aligned} &\text{Minimize} \quad \left(\frac{1}{\nu}-1\right)+W_2\frac{1}{\nu}(\frac{1}{\nu_2}-1)+W_2\frac{1}{\nu}\frac{1}{\nu_2}(\frac{1}{r_2^*}-1) \\ &\text{with respect to } \nu, \text{ and } \nu_2 \qquad (13) \\ &\text{Subject to} \quad (N c_0 + N c_1 W_1)\nu + N(c_2 + r_2^* c_3) W_2 \nu \, \nu_2 = C^* \end{aligned}$$

Our cost values are: $c_0 = 0.1$, $c_1 = 0.4$, $c_2 = 4.1$ and $c_3 = 0.4$, so that when $r_2^* = 1$ we obtain the cost of 4.5 as in Hansen and Hurwitz.

Table 2 shows the optimal values of ν and ν_2 for different values of C^* with $r_2^* = 1$ (complete response follow-up) and $r_2^* = 0.5$ (one-half of the follow-up sample responds).

Table 2: Optimal sampling fractions for different values of C^* comparing $r_2^* = 1$ with $r_2^* = 0.5$

	$r_2^* = 1$			$r_2^* = 0.5$		
C^*	ν	ν_2	Min Variance	ν	ν_2	Min Variance
1500	0.033	0.365	54.9	0.026	0.528	90.6
2000	0.045	0.365	40.9	0.035	0.528	67.7
2500	0.056	0.365	32.5	0.044	0.528	54.0
3000	0.067	0.365	27.0	0.052	0.528	44.8
3500	0.078	0.365	23.0	0.061	0.528	38.3
4000	0.089	0.365	20.0	0.070	0.528	33.4
4500	0.100	0.365	17.6	0.078	0.528	29.5
5000	0.111	0.365	15.8	0.087	0.528	26.5

Note: (14) agrees with equation (14) given in Rao (1973) when $r_2^* = 1$.

For the special case that $S_2^2 = S^2$, ν_2 is constant for each value of r_2^*. This value can be obtained by putting $L = 2$ in (10), which leads to the following expression.

$$\nu_2 = \sqrt{\frac{c_0 + c_1 W_1}{r_2^* W_1 (c_2 + r_2^* c_3)}} \tag{14}$$

The results are as follows. For a given r_2^*, increasing C^* implies that we only need to increase ν. On the other hand, for a given C^*, decreasing r_2^* implies that we should decrease ν and increase ν_2.

4. Bias considerations

Bethlehem (1988) observes that there are two different approaches for handling nonresponse. The first approach is to follow-up intensively a sample of the nonrespondents, and the second one is to reweight the data. We have discussed the first one in section 2. To discuss the second approach, we assume that the sample has been selected with simple random sampling without replacement. The extension to an arbitrary design is straightforward. Given that *n* units have been selected from a population of *N* units, assume that only $n_1 < n$ units respond to the survey. An estimator of the total $Y = \sum_{k \in U} y_k$ that strictly uses the responding units is

$$\hat{Y}_1 = \frac{N}{n_1} \sum_{k \in s_1} y_k \tag{15}$$

The bias associated with estimator (15) can be evaluated in one of two ways: the fixed response approach, (Cochran 1977); and the random response approach (Hartley 1946). In the fixed response approach, the population is split into a responding stratum U_1 with mean $\overline{Y}_1$ and a nonresponding stratum U_2 with mean $\overline{Y}_2$. The bias of estimator (15) depends on the difference between the means of those two strata, that is $B\left(\hat{Y}_1\right) = N_2\left(\overline{Y}_1 - \overline{Y}_2\right)$. Note that we can establish this result via a model-based approach. We suppose that the following model holds: $y_k = \mu_1 + e_k$ if $k \in U_1$ and $\mu_2 + e_k$ if $k \in U_2$, where $E_\zeta\left(e_k\right) = 0$ and $V_\zeta\left(e_k\right) = \sigma^2$. The model expectation of Y is $E_\zeta(Y) = \left(N_1\mu_1 + N_2\mu_2\right)$. The expectation of $\hat{Y}$ is $E_d E_\zeta\left(\hat{Y}_1\right) = N_1\mu_1$ where $E_d(\bullet)$ is the expectation under the model. Hence the bias associated with $\hat{\hat{Y}}_1$ is $N_2\left(\mu_1 - \mu_2\right)$.

In the random approach, each unit in the population has a response probability (unknown) θ_k, where $\theta_k > 0$ that are independent between units. Each unit in the population is assigned a value r_k which is equal to 1 if it responds and 0 otherwise: the expectation of

r_k given the response probability mechanism (say ξ**)** is given by $E_\xi\left(r_k\right)$. The approximate bias of $\hat{Y}_1$ is given by

$$B\left(\hat{Y}_1\right) \doteq N\frac{\sum_U y_k\ \theta_k}{\sum_U \theta_k} - Y \tag{16}$$

Note that the fixed approach can be viewed as an extreme case of the random approach. Units in the responding stratum U_1 respond with probability $\theta_k = 1$ while those in the non responding stratum U_2 "respond" with probability $\theta_k = 0$.

If auxiliary data ($\boldsymbol{x}_k$) are available, Lundström and Särndal (1999) suggested the following estimators (as applied to SRSWOR): $\hat{Y}_{1U} = (N/n)\sum\limits_{k\in s_1} g_{kU}\, y_k$ or $\hat{Y}_{1s} = (N/n)\sum\limits_{k\in s_1} g_{ks}\, y_k$. The g_{kU} and g_{ks} terms given by $1+\left(\sum\limits_{k\in U}\boldsymbol{x}_k - \sum\limits_{k\in s_1}\frac{N}{n}\boldsymbol{x}_k\right)\left(\sum\limits_{k\in s_1}\frac{N}{n}\boldsymbol{x}_k\boldsymbol{x}_k^T\right)^{-1}\boldsymbol{x}_k$ and $1+\left(\sum\limits_{k\in s}\boldsymbol{x}_k - \sum\limits_{k\in s_1}\frac{N}{n}\boldsymbol{x}_k\right)\left(\sum\limits_{k\in s_1}\frac{N}{n}\boldsymbol{x}_k\boldsymbol{x}_k^T\right)^{-1}\boldsymbol{x}_k$ are derived respectively from the constraints $\sum\limits_{k\in s_1} g_{kU}\,\boldsymbol{x}_k = \sum\limits_{k\in U}\boldsymbol{x}_k$ and $\sum\limits_{k\in s_1} g_{ks}\,\boldsymbol{x}_k = \sum\limits_{k\in s}\boldsymbol{x}_k$. Note that $\hat{Y}_{1U}$ requires the knowledge of $\boldsymbol{X} = \sum_U \boldsymbol{x}_k$ whereas $\hat{Y}_{1s}$ only assumes that $\boldsymbol{x}$ is not missing in the sample (e.g. poorer data).

Lundström and Särndal (1999) work out the bias of $\hat{Y}_{1s}$ and $\hat{Y}_{1U}$ as $-\sum_U\left(1-\theta_k\right)\left(y_k - \boldsymbol{x}_k^T\boldsymbol{B}_{U\theta}\right)$, where $\boldsymbol{B}_{U\theta} = \left(\sum_U \theta_k\boldsymbol{x}_k\boldsymbol{x}_k^T\right)^{-1}\left(\sum_U \theta_k\boldsymbol{x}_k y_k\right)$ is the regression vector. Lundström and Särndal (1999) showed that $\hat{Y}_{1s}$ and $\hat{Y}_{1U}$ would be approximately unbiased, provided that $\theta_k^{-1} = 1 + \boldsymbol{\lambda}^T\boldsymbol{x}_k$, where $\boldsymbol{\lambda}$ is a vector of constants. If $\boldsymbol{x}_k$ is a vector of indicator variables, referring to weighting classes, then one does not need to estimate θ_k provided that MCAR holds within each class.

Note that if we use the fixed response approach $\hat{Y}_{1U}$ is potentially biased (and so is $\hat{Y}_{1s}$). To see this, we suppose that $y_k = \boldsymbol{x}_k^T\,\boldsymbol{\beta}_1 + e_k$ if $k\in U_1$ and $y_k = \boldsymbol{x}_k^T\,\boldsymbol{\beta}_2 + e_k$ if $k\in U_2$, where $E_\zeta\left(e_k\right)=0$ and $V_\zeta\left(e_k\right)=\sigma^2$. Under this model, the model expectation of Y is $E_\zeta(Y) = \left(\sum_{U_1}\boldsymbol{x}_k^T\,\boldsymbol{\beta}_1 + \sum_{U_2}\boldsymbol{x}_k^T\,\boldsymbol{\beta}_2\right)$. The expectation of $\hat{Y}_{1U}$ is

$$E_\zeta\left(\hat{Y}_{1U}\right) = \sum_{k\in U}\boldsymbol{x}_k^T\,\boldsymbol{\beta}_1 \tag{17}$$

The model bias of $\hat{Y}_{wU}$ is therefore

$$B_{\zeta}\left(\hat{Y}_{1U}\right) = \sum_{U_2} \boldsymbol{x}_k^T \ \left(\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1\right) \tag{18}$$

In practice, the θ_k 's are unknown so they need to be estimated (see Kott and Chang 2010 for more details on how they are estimated). Denoting $\hat{\theta}_k$ as the estimate value of θ_k , an estimator that incorporates these estimated response probabilities as well as auxiliary data is

$$\hat{Y}_{1U\hat{\theta}} = (N/n)\sum_{k\in s_1} g_{k\theta} y_k / \hat{\theta}_k . \tag{19}$$

$\hat{Y}_{1U\hat{\theta}}$ can be viewed as a calibration estimator respecting the following constraint (in the case of SRSWOR) $\sum_{k\in s_1} g_{k\theta}\, \boldsymbol{x}_k = \sum_{k\in U} \boldsymbol{x}_k / \theta_k$. This estimator is reasonable as it has no bias; the variance estimator for this case is available in Kim and Kim (2007). Note that Lundström and Särndal (1999) mentioned this estimator but did not include it in their simulation.

The above results use auxiliary data when there is no follow-up. If there is a follow-up and the nonresponding stratum cannot be partitioned into groups (as it was in section 2), we can only draw a sub-sample s_3 of size m_2 is selected from s_2 .

Two estimators that do not use auxiliary information for calibration purposes are possible. One uses the estimated response probabilities $\hat{\theta}_k$ and the other partitions s_3 into *J* response groups, s_{3j} $\left(j=1,...,J\right)$: the corresponding responding subsamples are denoted as s_{4j} $\left(j=1,...,J\right)$.The estimator that uses the estimated response probabilities is

$$\hat{Y}_{3\hat{\theta}} = \frac{N}{n}\left(\sum_{k\in s_1} y_k + \frac{n_2}{m_2}\sum_{k\in s_4}\frac{y_k}{\hat{\theta}_k}\right) \tag{20}$$

while the one that uses response groups is

$$\hat{Y}_3 = \frac{N}{n}\left(\sum_{k\in s_1} y_k + \frac{n_2}{m_2}\sum_{j=1}^{J}\frac{m_{2j}}{b_{2j}}\sum_{k\in s_{4j}} y_k\right) \tag{21}$$

Variance estimators for (20) can be obtained by extending the one-phase solution given by Kim and Kim (2007) to two-phase sampling. The variance for (21) is obtained following the development given in section 2. Auxiliary data can be built into these estimators as well.

5. Conclusion

In this paper, we have provided the methodology for optimally following-up a sample of nonrespondents given a fixed budget. This budget incorporates the assumption that there will also be nonresponse for the follow-up sample of the nonrespondents. Given the fixed budget, the allocation assumed that response rates are available from previous experience for the initial contact with the sampled units, as well as for the propensity groups that for the nonresponding units in the sample. A numerical solution to this problem can be obtained through nonlinear programming methods. That is, we minimized a variance function subject to expected cost constraints that are both nonlinear functions of the unknown first–phase and second-phase sampling fractions. The algorithm was applied to the retail data example of Hansen and Hurvitz (1946) to check that it was correct.

In the event that the nonresponse portion of the sample could not be broken out into homogeneous propensity groups, we studied the bias properties of a number of possible estimators.

References

Bethlehem J.G. (1988). Reduction of Nonresponse Bias through Regression Estimation. *Journal of Official Statistics*, Vol. 4. No.3. 1988, 251-260.

Frayer, W. E. (1979). Multi-level sampling designs for resource inventories. USDA Forest Serv, Rocky Mt Forest and Range Exp Stn, RM Contract 16-747-CA.

Fuller, W. A., Loughin, M. M., and Baker, H. D. (1994). Regression Weighting for the 1987–88 National Food Consumption Survey. Survey Methodology, 20, 75–85.

Hansen, M.H. and Hurwitz, W.H. (1946). The Problem of Nonresponse in Sample Surveys. *Journal of the American Statistical Association*, 41, 517-529.

Hartley, H.O., (1946). Discussion of Paper by F. Yates. *Journal of the Royal Statistical Society*, 109, 37.

Jeyaratnam S., Bowden D. C., Graybill F. A., and Frayer W. E. (1984). Estimation in multiphase designs for stratification. Forest Sci., Vol. 30, No.2, 484-491.

Kim J. K. and Kim J. J. (2007). Nonresponse weighting adjustment using estimated response probability. *The Canadian Journal of Statistics*, Vol. 35, No. 4, 2007, 501–514.

Kott P.S. and Chang, T. (2010). Using Calibration Weighting to Adjust for Nonignorable Unit Nonresponse. *Journal of the American Statistical Association* , Vol. 105, No. 491, Theory and Methods. 1265-1275.

Little, R.J. (1986). Survey Nonresponse Adjustments, *International Statistical Review*, 54, 139-157.

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

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

Rao, J.N.K. (1975). Unbiased variance estimation for multistage designs. *Sankhyā, C 77,* 133-139.