2014

Model-assisted sample design of a first phase survey with two second-phase surveys

Robert Graham Clark

Model-assisted sample design of a first phase survey with two second-phase surveys

Abstract

Suppose there is a first phase sample s 1 of m units, which is a simple random sample without replacement (SRSWOR) from a population of size N . Membership of a subpopulation of interest (referred to as A) is collected from this rst phase sample. Let N A be the population size of the subpopulation and N B be the remaining population. Write a i equal to 1 for the subpopulation and 0 otherwise. The first phase survey also collects auxiliary variables z i which may include a i . Let m A and m B be the first phase sample sizes of subpopulation members and non-members respectively.

National Institute for Applied Statistics Research Australia

University of Wollongong

Working Paper

18-14

Model-Assisted Sample Design of a First Phase Survey with Two Second-Phase Surveys

Robert Graham Clark

National Institute for Applied Statistics Research Australia, University of Wollongong, Wollongong NSW 2522. Phone +61 2 4221 5435, Fax +61 2 4221 4845. Email: anica@uow.edu.au

MODEL-ASSISTED SAMPLE DESIGN OF A FIRST PHASE SURVEY

WITH TWO SECOND-PHASE SURVEYS

Robert Graham Clark[1]

1 Assumed Sample Design

Suppose there is a first phase sample s_1 of m units, which is a simple random sample without replacement (SRSWOR) from a population of size N. Membership of a subpopulation of interest (referred to as A) is collected from this first phase sample. Let N_A be the population size of the subpopulation and N_B be the remaining population. Write a_i equal to 1 for the subpopulation and 0 otherwise. The first phase survey also collects auxiliary variables $\boldsymbol{z_i}$ which may include a_i. Let m_A and m_B be the first phase sample sizes of subpopulation members and non-members respectively.

Two non-overlapping second phase samples are selected from s_1. Survey 1 is designed to estimate the population total of a variable y, using $\boldsymbol{z}$ as an auxiliary variable which is hopefully correlated with y. Survey 2 is designed to estimate the population total and subpopulation total of a variable u, with the first phase sample used to achieve oversampling of the subpopulation.

It is assumed that both Survey 1 and Survey 2 are stratified SRSWOR from s_1, stratified by subpopulation membership. Let n_A and n_B be the Survey 1 sample sizes from strata A and B. Let q_A and q_B be the Survey 2 sample sizes

[1]National Institute for Applied Statistics Research Australia, University of Wollongong. Email: rclark@uow.edu.au

from the two strata.

In practice, it may be desirable to stratify using $\boldsymbol{z}$, not just subpopulation membership. A simplified setup is assumed here, because the requirement to oversample the subpopulation in Survey 2 may be the dominant constraint. This is not explicitly required for Survey 1, but is still included, because the requirement for subpopulation sample will be a bottleneck for the overall design, and so undersampling of the subpopulation in Survey 1 should be considered.

2 Model and Framework

The variable of interest in the Survey 1 is denoted y_i for unit i, and the variable of interest in Survey 2 is u_i. The following model is assumed:

$$\left.\begin{array}{lcl} E_M\left[y_i\right] & = & \mu_y \\ var_M\left[y_i\right] & = & \sigma^2 \\ E_M\left[y_i|\boldsymbol{z_i}\right] & = & \boldsymbol{\beta^T z_i} \\ var_M\left[y_i|\boldsymbol{z_i}\right] & = & \gamma^2 \\ E_M\left[u_i|\boldsymbol{z_i}\right] & = & E_M\left[u_i\right] = \theta \\ var_M\left[u_i|\boldsymbol{z_i}\right] & = & var_M\left[u_i\right] = \psi^2 \\ E\left[a_i\right] = \phi & & \end{array}\right\} \tag{1}$$

Independence for distinct units is also assumed. We write $\gamma^2 = \left(1 - R^2\right)\sigma^2$ where R^2 is the R-squared of the regression of y_i on $\boldsymbol{z_i}$. The parameter ϕ is the expected proportion of the population who are in the subpopulation.

A major assumption in the model is that the first phase variables have no predictive power for the Survey 2 variable u_i. In practice, this will be approximately true for some Survey 2 variables but not others, so we are designing for the worst case. In contrast, $\boldsymbol{z_i}$ may have some predictive power for Survey 1 variables.

Another assumption is that there are no population auxiliary variables. This can easily be incorporated into the results of this note, by assuming two-phase GREG estimation, using both population-level and first-phase-level auxiliary variables. The only change is that y_i and u_i would be redefined as the residuals given the population auxiliary variables. This will be expressed more formally in a future paper.

The model-assisted framework will be used. Multi-phase GREG estimation is assumed. The anticipated variance (AV) is the model expectation of the design variance. In model-assisted sampling, sample designs are usually derived to optimise (at least approximately) the AV. This note will derive the AVs, and optimise them with respect to the design parameters m, n_A, n_B, q_A and q_B, subject to constraints. Section 2 sets up the model and framework. Section 3 defines the design problem and states the AVs. Section 4 contains the solution. The optimal design calculation has been implemented in R and a future version of this paper will include numerical results for various parameter values.

3 Defining the Objectives of the Design

Let C_1 be the cost per first phase unit. Let C_{2A} and C_{2B} be the cost per subpopulation and non-subpopulation Survey 1 selection, and C_{3A} and C_{3B} be the cost per subpopulation and non-subpopulation Survey 2 selection. The total cost is

$$C = C_1 m + C_{2A} n_A + C_{2B} n_B + C_{3A} q_A + C_{3B} q_B \tag{2}$$

The AV for the estimated totals of y is:

$$\begin{aligned}
AV\left[\hat{t}_y\right] &= \frac{N^2}{m}\sigma^2 + E\left[\frac{N^2}{m^2}\left(\frac{m_A^2}{n_A}\left(1-\frac{n_A}{m_A}\right)\gamma^2 + \frac{m_B^2}{n_A}\left(1-\frac{n_B}{m_B}\right)\gamma^2\right)\right] \\
&= \frac{N^2}{m}\sigma^2 + E\left[\frac{N^2}{m^2}\left(\frac{m_A^2}{n_A} - m_A + \frac{m_B^2}{n_B} - m_B\right)\sigma^2\left(1-R^2\right)\right] \\
&\approx \frac{N^2}{m}\sigma^2 - \frac{N^2}{m}\left(\frac{(\phi m)^2}{n_A} + \frac{((1-\phi)\,m)^2}{n_B}\right)\sigma^2\left(1-R^2\right) - \frac{N^2}{m}\sigma^2\left(1-R^2\right) \\
&= \frac{N^2}{m}\sigma^2 R^2 + N^2\left(\frac{\phi^2}{n_A} + \frac{(1-\phi)^2}{n_B}\right)\sigma^2\left(1-R^2\right) \qquad (3)
\end{aligned}$$

assuming that m/N is negligible. Details of these derivations are omitted but will be included in a future version of this paper. The derivation is broadly similar to the derivations for two-stage sampling for subpopulations in Clark (2009). Anticipated variances results for stratification can be found in Särndal et al. (1992), although this text does not have a result for the precise situation considered here. Similarly, we can obtain the AV for $\hat{t}_u$, noting also that the equivalent of R^2 for variable u is 0:

$$AV\left[\hat{t}_y\right] = N^2\left(\frac{\phi^2}{q_A} + \frac{(1-\phi)^2}{q_B}\right)\psi^2 \tag{4}$$

The AV for the estimate of the total of u for the subpopulation is:

$$AV\left[\hat{t}_{yA}\right] = N^2\phi^2 q_A^{-1}\psi^2 \tag{5}$$

The approach will be to minimise the total cost subject to constraints on the AVs of $\hat{t}_y$, $\hat{t}_u$ and $\hat{t}_{uA}$. Rather than setting values for the AVs, which would be difficult to interpret, we will suppose that the AV of $\hat{t}_y$ must be at least as good as the AV that would be achieved by a single phase SRSWOR of size k_y. This AV is equal to $N^2 k_y^{-1} \sigma^2$. Equating this to (3), we get the first constraint:

$$m^{-1} R^2 + n_A^{-1} \phi^2 \left(1 - R^2\right) + n_B^{-1} \left(1 - \phi\right)^2 \left(1 - R^2\right) \leq k_y^{-1} \tag{6}$$

Similarly, we require the effective sample size for estimating t_u to at least k_u, giving us the constraint

$$q_A^{-1} \phi^2 + q_B^{-1} \left(1 - \phi\right)^2 \leq k_u^{-1}. \tag{7}$$

We require the effective subpopulation sample size for estimating t_{uA} to be at least k_{uA}, giving us the constraint

$$q_A^{-1} \leq k_{uA}^{-1}. \tag{8}$$

There are also some inequality constraints, because the first phase sample in each stratum must be at least as large as the combined Survey 1 and Survey 2 sample sizes in each stratum:

$$n_A + q_A \leq m_A \approx m\phi \tag{9}$$

$$n_B + q_B \leq m_B \approx m\left(1 - \phi\right) \tag{10}$$

We will assume that (10) will be satisfied, since the non-subpopulation population is not to be heavily oversampled. Inequality (9) will be important, because under some scenarios the first phase subpopulation sample size will be a limiting factor.

So, the problem is to minimise the cost (2) with respect to m, n_A, n_B, q_A, q_B subject to constraints (6), (7) and (8).

4 Optimal Design Derivation

4.1 Initial Comments

Firstly, (7) must be satisfied with equality, since otherwise we can reduce the cost by reducing q_B without violating any constraints. Similarly, equality must hold in (8) at the optimum, otherwise we can reduce the cost by reducing q_A without violating any constraints. With equality, (7) and (8) immediately determine q_A and q_B:

The problem simplifies rapidly, as (8) immediately determines that $q_A = k_{uA}$, and (7) then gives q_B:

$$q_A = k_{uA} \tag{11}$$

$$q_B = \left\{k_u^{-1} - q_A^{-1}\phi^2\right\}^{-1}(1-\phi)^2 \tag{12}$$

Constraint (6) must also hold with equality, because otherwise we can reduce the cost with impunity by reducing either n_A or n_B.

The problem is now to minimise the cost (2) with respect to m, n_A and n_B subject to (6) with equality and to (10).

There are two possibilities depending on whether constraint (10) is active or inactive.

4.2 First Case: (10) is Inactive

The first possibility is that (10) is inactive, i.e. if we optimise ignoring the constraint, it turns out to be satisfied anyway. Ignoring (10), this is a standard Neymann allocation problem. The solution can be derived either using the Cauchy-Schwarz inequality or Lagrange multipliers (e.g. Clark and Steel 2000), and is:

$$\left.\begin{array}{lll} m & = & \lambda\sqrt{R^2/C_1} \\ n_A & = & \lambda\phi\sqrt{\left(1-R^2\right)/C_{2A}} \\ n_B & = & \lambda\left(1-\phi\right)\sqrt{\left(1-R^2\right)/C_{2B}} \\ \text{where} & & \\ \lambda & = & k_y\left\{\sqrt{C_1R^2}+\phi\sqrt{C_{2A}\left(1-R^2\right)}+\left(1-\phi\right)\sqrt{C_{2B}\left(1-R^2\right)}\right\}. \end{array}\right\} \tag{13}$$

For simplicity, it will be assumed from here on that $C_{2A} = C_{2B} = C_2$. With this simplification, (13) becomes

$$\left.\begin{array}{lll} m & = & \lambda\sqrt{R^2/C_1} \\ n_A & = & \lambda\phi\sqrt{\left(1-R^2\right)/C_2} \\ n_B & = & \lambda\left(1-\phi\right)\sqrt{\left(1-R^2\right)/C_2} \\ \lambda & = & k_y\left\{\sqrt{C_1R^2}+\sqrt{C_2\left(1-R^2\right)}\right\}. \end{array}\right\} \tag{14}$$

We can also obtain the second phase sampling fractions $f_A = n_A/m_A \approx n_A/\left(m\phi\right)$ and $f_B = n_B/m_B \approx n_B/\left(m\left(1-\phi\right)\right)$ at this optimum:

$$f_A = f_B = \sqrt{\frac{C_2}{C_1}\frac{\left(1-R^2\right)}{R^2}} \tag{15}$$

So we have equal probability sampling over both subpopulation and non-subpopulation members.

We can check numerically for any given set of values of C_1/C_2, R^2, ϕ, k_y and k_{uA} whether (14) satisfies (10) or not. If it does, then (14) is the optimum. If it doesn't, then the constraint is active, and we must go on to the next heading.

4.3 Second Case: (10) is Active

In this case, equality holds in (10), so that m is determined by n_A and q_A:

$$m = (n_A + q_A)\,\phi^{-1}.$$

The problem then becomes to minimise the cost

$$\begin{aligned} C &= C_1\,(n_A + q_A)\,\phi^{-1} + C_2 n_A + C_2 n_B + C_{3A} q_A + C_{3B} q_B \\ &= \left(C_1 \phi^{-1} + C_2\right) n_A + C_2 n_B + \mathit{constants} \qquad (16) \end{aligned}$$

with respect to n_A and n_B, where "constants" refers to terms that do not depend on n_A or n_B, subject to constraint (6) which becomes

$$(n_A + q_A)^{-1}\,\phi R^2 + n_A^{-1}\phi^2\left(1 - R^2\right) + n_B^{-1}\,(1-\phi)^2\left(1 - R^2\right) = k_y^{-1} \qquad (17)$$

Notice that the cost coefficient attached to n_A in (16) is now $C_1\phi^{-1} + C_2$ rather than C_2. This reflects the fact that when (10) is active, we must screen an additional ϕ^{-1} households for every additional subpopulation unit that we want in Survey 1. In contrast, when (10) is inactive, the value of m which is optimal for Survey 1 is large enough that there are more subpopulation units than we need.

Now we strike a problem, because minimising (16) subject to (17) is no longer a Neymann allocation problem. Using Lagrange multipliers, we get:

$$
\begin{aligned}
0 &= C_1\phi^{-1} + C_2 - \lambda\left(n_A + q_A\right)^{-2}\phi R^2 - n_A^{-2}\phi^2\left(1 - R^2\right) \\
0 &= C_2 - \lambda n_B^{-1}\left(1 - \phi\right)^2\left(1 - R^2\right)
\end{aligned}
$$

Solving this with respect to n_A and n_B turns out to involve a quartic equation. So an analytic solution is not feasible.

Instead, the problem can be solved numerically, by setting up the cost as a function of n_A which is then minimised. For each value of n_A, the value of n_B is obtained by numerically solving the constraint (17). Then the cost can be calculated using (16).

4.4 Adding a Constraint of Equal Probability in Survey 2 when (10) is Active

When (10) is inactive, we saw in Section 4.2 that $f_A = f_B$. When the constraint is active, this turns out not to be the case, as seen in 4.3. What if we decide that undersampling of subpopulation members is not acceptable in Survey 1? This leads to an additional constraint, that $n_A/\phi = n_B/\left(1 - \phi\right)$. Constraint (17) becomes:

$$
\begin{aligned}
k_y^{-1} &= \left(n_A + q_A\right)^{-1}\phi R^2 + n_A^{-1}\phi^2\left(1 - R^2\right) + \left(n_A\left(1 - \phi\right)\phi^{-1}\right)^{-1}\left(1 - \phi\right)^2\left(1 - R^2\right) \\
&= \left(n_A + q_A\right)^{-1}\phi R^2 + n_A^{-1}\phi^2\left(1 - R^2\right) + n_A^{-1}\phi\left(1 - \phi\right)\left(1 - R^2\right) \\
&= \left(n_A + q_A\right)^{-1}\phi R^2 + n_A^{-1}\phi^2\left(1 - R^2\right) + n_A^{-1}\left(\phi - \phi^2\right)\left(1 - R^2\right) \\
&= \left(n_A + q_A\right)^{-1}\phi R^2 + n_A^{-1}\phi\left(1 - R^2\right) \qquad (18)
\end{aligned}
$$

This can be re-expressed as a quadratic equation in n_A. It can be solved numerically or analytically. n_A and m follow immediately.

References

Clark, R. G. (2009), “Sampling of subpopulations in two-stage surveys,” *Statistics in Medicine*, 28, 3697–3717.

Clark, R. G. and Steel, D. G. (2000), “Optimum Allocation of Sample to Strata and Stages with Simple Additional Constraints,” *Journal of the Royal Statistical Society, Series D: The Statistician*, 49, 197–207.

Särndal, C., Swensson, B., and Wretman, J. (1992), *Model Assisted Survey Sampling*, New York: Springer-Verlag.