Centre for Statistical and Survey Methodology

The University of Wollongong

Working Paper

02-12

Potential Gains from Sample Design Using Unit Level Cost Information

David Steel

Centre for Statistical & Survey Methodology, University of Wollongong, Wollongong NSW 2522. Phone +61 2 4221 5435, Fax +61 2 4221 4845. Email: anica@uow.edu.au

Potential Gains from Sample Design Using Unit Level Cost Information

David Steel

Centre for Statistical and Survey Methodology

University of Wollongong

Abstract

In developing the sample design for a survey we usually attempt to produce a good design for the funds available. Information on costs can be used to develop sample designs that minimise the sampling variance of an estimator of total for fixed costs. Improvements in survey management systems mean that it is now possible to estimate the cost of including each unit in the sample. This paper develops relatively simple approaches to determine whether the potential gains arising from using this unit level cost information are likely to be of practical use. It is shown that the key factor is coefficient of variation of the costs relative to the coefficient of variation of the relative error on the estimated cost coefficients.

Key words: optimal design, sampling variance**,**

1. Introduction

In developing the sample design for a survey we usually attempt to produce a good design for the funds available. In many cases the sample size is taken as fixed and the design is developed to minimise the sampling variance of an estimator of mean or total. This approach assumes that the cost of enumerating a unit is the same for all

units in the population. If costs vary between units then this variation can be taken into account in the sample design. Information on costs can be used to develop sample designs that minimise the sampling variance of an estimator of total for fixed costs. In stratified sampling a common approach is to estimate a cost coefficient for each stratum and determine the optimal allocation. The resulting allocation of sample to strata is proportional to the inverse of the square root of the stratum cost coefficients (Neyman, 1934). In a multistage design the costs of including the units at the different stages of selection can be used to decide the number of units to select at each stage (Hansen, Hurwitz and Madow, 1953). Discussions of costs in survey design are given in Cochran (1977), Kish (1965) and Groves (1989). Clark and Steel (2000) summarize the key results.

In these approaches the costs at each stage are assumed to be constant within strata. In practice costs will vary across units and the cost coefficients used are essentially averages. Improvements in survey management systems mean that it is now possible to estimate the cost of including each unit in the sample. This paper develops relatively simple approaches to determine whether the potential gains arising from using this unit level cost information are likely to be of practical use.

In section 2 the gain in using the individual level cost information is considered for the simple case of Poisson sampling assuming that the cost and variate values are known and positive for each unit. The more realistic situation where the cost coefficients are estimated with some error and instead of the values of the variable of interest we know the values of an auxiliary variable are considered. The extension to cases where the population units may take zero values is also considered in section 3.

The case when the selection probabilities are changed at some point in the conduct of the survey because of a change in the budget is also considered in section 4. Section 5 gives a brief summary and discussion.

2. Poisson Sampling

2.1 Known Variate Values and Costs

Consider a finite population, U, consisting of values $Y_i, i \in U$. A sample is to be selected using a Poisson sampling scheme in which the ith population unit has probability of selection π_i. Let δ_i be an indicator for sample membership, so $P(\delta_i = 1) = \pi_i$. The cost of enumerating unit i is C_i and the total funds available is C_A, so that the cost of enumerating the sample s is $\sum_{i \in s} C_i$ and the expected cost is $\sum_{i \in U} C_i \pi_i$. The expected sample size is $n_E = \sum_{i \in U} \pi_i$. Assume that $\hat{T}_Y = \sum_{i \in s} \frac{Y_i}{\pi_i} = \sum_{i \in U} \delta_i \frac{Y_i}{\pi_i}$ is used to estimate $T_Y = \sum_{i \in U} Y_i$. The estimator $\hat{T}_Y$ is design unbiased for T_Y and has sampling variance

$$V(\hat{T}_Y) = \sum_{i \in U} \frac{Y_i^2}{\pi_i} - \sum_{i \in U} Y_i^2 \qquad (1)$$

For the moment will allow π_i to depend on Y_i. The more realistic case when we do not know the values of the variable of interest, but we do know the values of a related auxiliary variable is considered in section 2.2.

Theorem 1.1

If $Y_i > 0$ for all $i \in U$ then $V(\hat{T}_Y)$ is minimized for expected cost equal to C_A if $\pi_i = \frac{Y_i}{\sqrt{C_i}} \frac{C_A}{\sum_{i \in U} Y_i \sqrt{C_i}}$ and the resulting sampling variance is

$$V_{opt}(\hat{T}_Y) = \frac{\left(\sum_{i \in U} Y_i \sqrt{C_i}\right)^2}{C_A} - \sum_{i \in U} Y_i^2 \tag{1}$$

Proof: apply standard Lagrangian methods.

A useful result is given by

Lemma 1.2

Consider the pairs $u_i, v_{i,} i = 1,...N$, and the covariance $S_{uv} = \frac{1}{N}\sum_{i=1}^{N}(u_i - \bar{u})(v_i - \bar{v})$ then

$$\sum_{i=1}^{N} u_i v_i = N\bar{u}\bar{v}(1 + C_{u,v}) \tag{2}$$

where $C_{u,v} = \frac{S_{uv}}{\bar{u}\bar{v}}$ is the relative covariance of the values.

Setting $u_i = v_i$ gives

$$\sum_{i=1}^{N} u_i^2 = N\bar{u}^2(1 + C_u^2) \tag{3}$$

where C_u is the coefficient of variation of the values of u .

If $R_{u,v}$ is the correlation of the values of u and v then

$$\sum_{i=1}^{N} u_i v_i = N\bar{u}\bar{v}(1 + R_{u,v}C_u C_v) \tag{4}$$

Applying these results we obtain

Theorem 1.3

$$V_{opt}(\hat{T}_Y) = N^2\bar{Y}^2\left[\frac{\bar{\sqrt{C}}^2}{\bar{C}_A}\frac{\left(1 + C_{Y,\sqrt{c}}\right)^2}{n_E} - \frac{(1 + C_Y^2)}{N}\right] \tag{5}$$

where $\bar{\sqrt{C}} = \frac{1}{N}\sum_{i=1}^{N}\sqrt{C_i}$ is the population average of the square root of the cost coefficients and $\overline{C}_A = \frac{C_A}{n_E}$ is the average of the allocated cost per expected sample unit.

We will compare the optimal strategy with three approaches for the same expected cost:

1. ignoring costs
2. ignoring variate values
3. ignoring costs and variate values

Ignoring Costs

If the costs are ignored then $\pi_i \propto Y_i$. Fixing the expected cost at C_A gives $\pi_i = Y_i \frac{C_A}{\sum_{i\in U} Y_i C_i}$ and the resulting sampling variance $V_{Y.}(\hat{T}_Y) = \frac{\left(\sum_{i\in U} Y_i\right)\left(\sum_{i\in U} Y_i C_i\right)}{C_A} - \sum_{i\in U} Y_i^2$. Applying (2) gives

$$V_{Y.}(\hat{T}_Y) = N^2\overline{Y}^2\left[\frac{\overline{C}}{\overline{C}_A}\frac{\left(1+C_{Y,C}\right)}{n_E} - \frac{(1+C_Y^2)}{N}\right] \qquad (6)$$

where $\overline{C} = \frac{1}{N}\sum_{i\in U} C_i$ is the average of the cost coefficients.

Noting that $\overline{C} = \frac{1}{N}\sum_{i\in U}\left(\sqrt{C_i}\right)^2$ and applying (3) gives $\overline{C} = \left(\bar{\sqrt{C}}\right)^2(1+C^2_{\sqrt{C}})$, where $C^2_{\sqrt{C}}$ is the coefficient of variation of $\sqrt{C_i}$, and so

$$V_{Y.}(\hat{T}_Y) = N^2\overline{Y}^2\left[\frac{\left(\bar{\sqrt{C}}\right)^2}{\overline{C}_A}\frac{(1+C^2_{\sqrt{C}})}{n_E}\left(1+C_{Y,C}\right)-\frac{(1+C_Y^2)}{N}\right] \tag{7}$$

If the sampling fraction is not large then the last term in (5) and (7) can be ignored and so

$$\frac{V_{Y.}}{V_{opt}} \approx \frac{(1+C^2_{\sqrt{C}})(1+C_{Y,C})}{(1+C_{Y\sqrt{C}})^2} \tag{8}$$

If the costs and variate values are unrelated this gives

$$\frac{V_{Y.}}{V_{opt}} \approx (1+C^2_{\sqrt{C}}) \tag{9}$$

Ignoring Values

Suppose that only the cost are taken into account, so that $\pi_i \propto \frac{1}{\sqrt{C_i}}$, which for expected cost C_A implies $\pi_i = \frac{1}{\sqrt{C_i}}\frac{C_A}{\sum\sqrt{C_i}}$ and associated sampling variance

$$V_{.C}(\hat{T}_Y) = \frac{\left(\sum_{i\in U} Y_i^2\sqrt{C_i}\right)\left(\sum_{i\in U}\sqrt{C_i}\right)}{C_A} - \sum_{i\in U} Y_i^2 .$$

Applying (3) and (2)

$$V_{.C}(\hat{T}_Y) = N^2\overline{Y}^2\left[\frac{\left(\bar{\sqrt{C}}\right)^2}{\overline{C}_A}\frac{(1+C_Y^2)}{n_E}\left(1+C_{Y^2,\sqrt{C}}\right)-\frac{(1+C_Y^2)}{N}\right] \tag{10}$$

For small sampling fraction

$$\frac{V_{.C}}{V_{opt}} \approx (1+C_Y^2)\left(1+C_{Y^2,\sqrt{C}}\right) \tag{11}$$

If the costs and variate values are unrelated this gives

$$\frac{V_{.C}}{V_{opt}} \approx (1 + C_Y^2) \tag{12}$$

Ignoring Costs and Values

When cost and variate values are ignored the selection probabilities will be constant, which for fixed expected costs gives, $\pi_i = \dfrac{C_A}{\sum_{i \in U} C_i}$ and associated sampling variance

$V_{..}(\hat{T}_Y) = \dfrac{\left(\sum_{i \in U} Y_i^2\right)\left(\sum_{i \in U} C_i\right)}{C_A} - \sum_{i \in U} Y_i^2$. Applying (3) and (2)

$$V_{.C}(\hat{T}_Y) = N^2 \overline{Y}^2 \left[\frac{\left(\bar{\sqrt{C}}\right)^2}{\overline{C}_A} \frac{(1 + C_Y^2)}{n_E} \left(1 + C_{\sqrt{C}}^2\right) - \frac{(1 + C_Y^2)}{N} \right] \tag{13}$$

For small sampling fraction

$$\frac{V_{..}}{V_{opt}} \approx \frac{(1 + C_Y^2)\left(1 + C_{\sqrt{C}}^2\right)}{1 + C_{Y\sqrt{C}}} \tag{14}$$

If the costs and variate values are unrelated this gives

$$\frac{V_{..}}{V_{opt}} \approx (1 + C_Y^2)\left(1 + C_{\sqrt{C}}^2\right) \tag{15}$$

These results show that taking costs into account in the design eliminates the $\left(1 + C_{\sqrt{C}}^2\right)$term and taking the values into account the term $(1 + C_Y^2)$ is eliminated.

2.2 Using Auxiliary Variable and Estimated Costs

The analysis so far has assumed that we know the population values of the variable of interest and the cost coefficients precisely. In practice we will not know the variate values but we may know the values of a related auxiliary variable, X_i and the cost

coefficient C_i will be estimated, with some error, as D_i. Write $Y_i = a_i X_i$ and $D_i = b_i C_i$.

Using the auxiliary variable and the estimated costs in the optimal probabilities implies $\pi_i \propto \frac{X_i}{\sqrt{D_i}}$. To make comparisons for the same true expected costs we consider $\pi_i = \frac{X_i}{\sqrt{D_i}} \frac{C_A}{\sum_{i \in U} C_i X_i / \sqrt{D_i}}$. The resulting sampling variance is $V_{XD}(\hat{T}_Y) = \frac{\left(\sum_{i \in U} Y_i^2 \sqrt{D_i} / X_i\right)\left(\sum_{i \in U} C_i X_i / \sqrt{D_i}\right)}{C_A} - \sum_{i \in U} Y_i^2$. This can be expressed in terms of the relative error factors a_i and b_i, as

$$V_{XD}(\hat{T}_Y) = \frac{\left(\sum_{i \in U} a_i^2 X_i \sqrt{D_i}\right)\left(\sum_{i \in U} \sqrt{C}_i X_i / \sqrt{b_i}\right)}{C_A} - \sum_{i \in U} Y_i^2 .$$

Theorem 1.4

$V_{XD}(\hat{T}_Y) = N^2 \overline{Y}^2 \left[\frac{\bar{\sqrt{C}}^2}{\overline{C}_A} \frac{\left(1 + C_a^2\right)^2 (1 + C_{\sqrt{b}}^2)}{n_E} A - \frac{(1 + C_Y^2)}{N} \right]$, where

$A = (1 + C_{X,\sqrt{D}})(1 + C_{a^2,X\sqrt{D}})(1 + C_{X,\frac{1}{\sqrt{b}}})(1 + C_{\sqrt{C},X/\sqrt{b}})(1 + C_{\sqrt{b},\sqrt{C}})(1 + C_{aX})^{-2}$

Proof

Apply (2) and (3) to sums of products and note that (3) can be used to obtain $N^2 \overline{Y}^2 = N^2 \overline{a}^2 \overline{X}^2 (1 + C_{aX})^2$ and $\bar{\sqrt{D}} = \bar{\sqrt{b}}\,\bar{\sqrt{C}}(1 + C_{\sqrt{b},\sqrt{C}})$. Taylor series methods give $\bar{\sqrt{b}}\left(\frac{1}{N}\sum_{i \in U} b_i\right) \approx (1 + C_{\sqrt{b}}^2)$.

Assuming that the relative error factors are independent of the costs and values and also that the estimated costs are independent of the auxiliary variable, then $A = 1$. In this case

$$V_{XD}(\hat{T}_Y) = N^2\overline{Y}^2\left[\frac{\bar{\sqrt{C}}^2}{\overline{C}_A}\frac{\left(1+C_a^2\right)^2(1+C_{\sqrt{b}}^2)}{n_E} - \frac{(1+C_Y^2)}{N}\right] \tag{16}$$

For small sampling fractions, when $A = 1$, we obtain

$$\frac{V_{XD}}{V_{opt}} \approx (1+C_a^2)(1+C_{\sqrt{b}}^2) \tag{17}$$

This shows the effect of the use of the auxiliary variable and the estimated cost coefficients is determined by the coefficient of variation of a_i and $\sqrt{b_i}$. Comparing this variance with completely ignoring the auxiliary variable and the costs, gives

$$\frac{V_{XD}}{V_{..}} \approx \frac{(1+C_a^2)(1+C_{\sqrt{b}}^2)}{(1+C_Y^2)(1+C_{\sqrt{C}}^2)} \tag{18}$$

When considering the gains from using the costs information we should consider what would happen when we use the auxiliary variable but ignore the estimated costs. Using similar methods we get the associated sampling variance as

$$V_{X.}(\hat{T}_Y) = N^2\overline{Y}^2\left[\frac{\bar{\sqrt{C}}^2}{\overline{C}_A}\frac{(1+C_{\sqrt{C}}^2)(1+C_a^2)(1+C_{a^2,X})(1+C_{C,X})}{n_E} - \frac{(1+C_Y^2)}{N}\right] \tag{19}$$

When independence assumptions are made this becomes

$$V_{X.}(\hat{T}_Y) = N^2\overline{Y}^2\left[\frac{\bar{\sqrt{C}}^2}{\overline{C}_A}\frac{(1+C_{\sqrt{C}}^2)(1+C_a^2)}{n_E} - \frac{(1+C_Y^2)}{N}\right] \tag{20}$$

For small sampling fraction the increase in variance arising from ignoring the estimated cost coefficients is

$$\frac{V_{X.}}{V_{XD}} = \frac{(1 + C^2_{\sqrt{C}})}{(1 + C^2_{\sqrt{b}})} \tag{21}$$

Hence, provided the coefficient of variation of the square root of the relative errors in estimating the cost is less than the coefficient of variation of the square root of the costs, there is a gain in using the information on individual level costs. There is a gain in using the auxiliary variable, provided the coefficient of variation of the factor relating the variable of interest to the auxiliary variable is less than the coefficient of variation of the variable of interest.

Example

Suppose $\frac{C_i}{\overline{C}}$ mainly varies between 0.25 and 4, so that $\frac{\sqrt{C_i}}{\sqrt{\overline{C}}}$ varies approximately uniformly between 0.5 and 2, then $C_{\sqrt{C}} \approx 0.4$ and from (9) $\frac{V_{Y.}}{V_{opt}} \approx 1.16$. Thus the loss through ignoring the costs is 16 percent and correspondingly the gain in using them is 15 percent. Similarly if $\frac{C_i}{\overline{C}}$ mainly varies between 0.2 and 2, then $C_{\sqrt{C}} \approx 0.031$ and $\frac{V_{Y.}}{V_{opt}} \approx 1.031$, and a potential gain of about 3 percent might arise from using the cost information.

Suppose $\frac{a_i}{\overline{a}}$ mainly varies approximately uniformly between 0.6 and 1.4, so $C_a \approx 0.2$ and $(1 + C_a^2) = 1.04$. From (17) this suggests that there will be a 4 percent loss through using the auxiliary variable instead of the values of the variable on interest.

Suppose $\frac{b_i}{\bar{b}}$ mainly varies between 0.6 and 1.4, so that $\frac{\sqrt{b_i}}{\sqrt{\bar{b}}}$ varies between 0.8 and 1.2, then $C_{\sqrt{b}} \approx 0.1$ and so $(1 + C^2_{\sqrt{b}}) = 1.01$. Assuming $C_{\sqrt{C}} \approx 0.4$ as above, then from (21) $\frac{V_{X.}}{V_{XD}} = \frac{1.16}{1.01} = 1.15$. Hence the use of estimated rather that actual costs reduces the gain by 1 percent, but a gain of 15% still arises from the use of the cost information.

The simple results are based on various covariances being zero. While this may not always be true, the covariances will often be small. Further empirical and theoretical work could be done to check this out. For many cases there would be no reason for there to be an appreciable covariance. Small covariances between the relative errors and the variable that they apply to depend on errors behaving in this way. The most likely case where the assumption that the covariance is small may be questionable is the covariance between cost and the variate value. In some cases, larger values of Y_i may lead to larger costs. A particular case is that of a dichotomous variable, which is considered in section 4.

Even though simple Poisson sampling is not used in practice, the theory described here could be used to calculate relative efficiencies. The theory shows that we need to have some information on the costs, the relationship between the estimated and actual costs, and between the variable of interest and the auxiliary variable. ABS data could be used to examine these relationships. Relative efficiencies could be calculated without making the assumptions concerning the covariances.

The theory can be expanded to consider the case when the population size is known and we use the estimator $\tilde{T}_Y = \frac{N}{\sum_{i \in s} \pi_i^{-1}} \hat{T}_Y$.

3 Allowing for Variables Taking Zero Values

The analysis so far does not allow for the case when Y_i can take the value 0, which means that it does not apply for a dichotomous variable. Also, we directly used the auxiliary variable in determining π_i and it is not clear that this is the best way of using the auxiliary variable. Both these issues can be tackled by assuming that there is some statistical model that relates Y_i to the population values of the auxiliary variables, $\mathbf{X}_U$. Assume that the selection probabilities are determined by $\mathbf{X}_U$. Then, for Poisson sampling the expectation of the sampling variance is

$$V(\hat{T}_Y \mid \mathbf{X}_U) = \sum_{i \in U} \frac{E[Y_i^2 \mid \mathbf{X}_U]}{\pi_i} - \sum_{i \in U} E[Y_i^2 \mid \mathbf{X}_U]$$

As the estimator is design unbiased this is also the total variance. Hence, all the results obtained above apply with Y_i replaced by ϕ_i, where $\phi_i^2 = E[Y_i^2 \mid \mathbf{X}_U] = E[Y_i \mid \mathbf{X}_U]^2 + V(Y_i \mid \mathbf{X}_U)$. Also, if we estimate ϕ_i by $\hat{\phi}_i$ then $a_i = \frac{\phi_i}{\hat{\phi}_i}$.

For a dichotomous variable indicating membership of a particular category of the population, for example being unemployed, $Y_i^2 = Y_i$ and so $\phi_i = E[Y_i \mid \mathbf{X}_U]^{0.5} = P(Y_i = 1 \mid \mathbf{X}_U)^{0.5}$.

If $\mathbf{X}_U$ indicates geographic areas, such as regions, then ϕ_i reflects the square root of the probability of people being in the category for the region. Often the variation of

ϕ_i is small and so there is little loss in ignoring it in determining the selection probabilities.

4 Adjustment of Design During the Enumeration Period

In some cases the selection probabilities may be reviewed as the survey progresses and adjustments made to the selection probabilities in light of the funds spent or changes made to the budget. To approximate this situation suppose the population is divided into two components, U_1 and U_2. Suppose that the units in U_1 have been subjected to the Poisson selection scheme using selection probabilities π_i resulting in the sample s_i and the selections are represented by δ_i. The selection probabilities used for U_1 are not necessarily chosen optimally. At this stage the funds still available are $C_{A2} = C_A - \sum_{i \in U_1} C_i \delta_i$. The sampling variance will be $V(\hat{T}_Y) = \sum_{i \in U_1} \frac{Y_i^2}{\pi_i} + \sum_{i \in U_2} \frac{Y_i^2}{\pi_i} - \sum_{i \in U} Y_i^2$. At this stage the only term in the variance that can be affected is the middle term, thus the optimal choice of the selection probabilities for U_2 are given by $\pi_i = \frac{Y_i}{\sqrt{C_i}} \frac{C_{A2}}{\sum_{i \in U_2} Y_i \sqrt{C_i}}$ for $i \in U_2$.

5. Summary and Conclusions

The theoretical results developed here show that for there to be appreciable gain from exploiting information on the costs of enumerating units in a sample survey the square root of these costs must vary considerably. While derived for the simple case of Poisson sampling the formulas can be used to quickly judge the potential gain from taking costs into account, by assessing the coefficient of variation of the square root of

the costs. Costs will be estimated with some error and this reduces the gain from using cost information by a factor determined by the relative variation of the square root of the relative errors in estimating the costs at the individual level.

References

Clark, R. and Steel, D. G. (2000). Optimum allocation to strata and stages with simple additional constraints. *The Statistician,* l**49**, 197-207.

Cochran, W. G. (1977). *Sampling Techniques*, 3rd edn. New York: Wiley.

Groves, R, (1989). *Survey Errors and Survey Costs*. New York: Wiley.

Hansen, M. H., Hurwitz, W. N. and Madow, W. G. (1953). *Sample Survey Methods and Theory*. New York: Wiley.

Kish. L. (1965). *Survey Sampling*. New York: Wiley.

Neyman, J. (1934). On two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society* **97**, 558-606