TWO-STAGE SAMPLE DESIGN WITH SMALL CLUSTERS

Robert G. Clark and David G. Steel
School of Mathematics and Applied Statistics, University of Wollongong, NSW 2522 Australia.
(robert.clark@abs.gov.au)

Key Words: sample design, household surveys, telephone surveys.

Abstract:

In two-stage surveys, the sample size of clusters and units within clusters can be chosen to minimise the variance of an estimator, for fixed cost. This paper considers sample designs where the number of units to be selected from each cluster is a function of the cluster size. If there are only a small number of units in each cluster, as in household surveys, then the optimisation should be over integers. An integer programming method is developed which gives significantly lower variances than traditional methods. A non-integer within-cluster sample size can be implemented by using a mixture of several integers; this can further reduce the variance.

1. Introduction

In two-stage surveys, a sample of clusters is selected, followed by a sample of units from each selected cluster. There are several possible reasons for this approach. There may be a list of the clusters in the population, but not of the units: for example, there is rarely a list of the people in the general population but there may be a list of households either for the whole population or within particular areas by a field listing exercise. Two stage surveys are also useful so that the sample can be made more geographically clustered, which often reduces the enumeration cost.

This article assumes that clusters are selected by a simple random sample without replacement (SRSWOR) of size m. Each cluster g contains N_g units. A SRSWOR of n_g units is selected from each selected cluster. The population and sample of clusters will be denoted U^1 and s^1 respectively. The values of n_g are assumed not to depend on s^1.

The sample design problem is to choose m and n_g. The allocation of sample to the first and second stages is a well-known problem (e.g. Hansen et al., 1953, ch.6,ch.7). The allocation of cluster and unit sample sizes is a balance between costs and variances. Cost typically consists of a per-cluster component (for example travel time if clusters are households or geographic areas) and a per-unit component (for example interview and processing time). If the per-cluster costs are much higher than the per-unit costs, then it is appropriate to select a highly clustered sample; that is the number of units selected in each selected cluster should be high. However, a highly clustered sample has higher variance if there are positive correlations between the values of units in the same cluster.

For some designs, the optimal sample sizes n_g are proportional to $N_g S_g$ where S_g^2 is the adjusted population variance for cluster g (Hansen et al., 1953, ch.6,ch.7). In practice there is not usually information to estimate S_g^2 for every separate cluster g, and the values of S_g^2 may not vary much between clusters. As a result, choosing n_g proportional to N_g will often give a reasonably efficient design.

This paper considers sample designs where N_g (and hence n_g) are small integers, so that the real-valued optima derived by Hansen et al. (1953) and others may not be the best designs possible. It is assumed that n_g are a function of N_g, say $n_g = \bar{n}_a$ for $N_g = a$. The problem is to choose m and $\bar{n}_a$ for $a = 1, \ldots, A$ where A is the largest cluster size.

A common example of this is household sampling, where clusters are households and units are people. In practice, either all people or one randomly selected person are usually surveyed; this article will suggest some more efficient alternatives.

The expected cost of implementing the survey is assumed to be

$$C = C_0 + C_1 m + C_2 \sum_{a=1}^{A} n_a \qquad (1)$$

where: $n_a = \frac{M_a}{M} m \bar{n}_a$ is the expected sample size of units in clusters of size a; M_a is the number of clusters of size a in the population; C_0 is fixed costs; C_1 is for costs associated with the number of clusters in the sample (for example travel costs); and C_2 is for costs associated with the number of units in the sample (for example interview time).

Suppose that the mean, variance and intraclass

This research was jointly supported by the Australian Research Council and the Australian Bureau of Statistics.

correlation of the variable of interest do not depend on the cluster size. These assumptions are made for simplicity; (Clark, 2002) gives more general results. The variance of an inverse selection probability estimator is then proportional to

$$V = R\frac{M^2}{m} + (1-R)\sum_{a=1}^{A}\frac{N_a^2}{n_a} \tag{2}$$

where: M is the number of clusters in the population; N_a is the number of units in the population in clusters of size a; and R is the finite population intracluster correlation coefficient.

In Section 2, the optimal values of m and $\bar{n}_a$, which minimize the variance for fixed cost, are discussed. The standard optimal allocation, which ignores the fact the m and $\bar{n}_a$ are integers, is stated. An algorithm which finds the best integer values of $\bar{n}_a$ is derived. In Section 3, clusters of size a are randomly assigned integer sample sizes, so that the expected sample size within clusters of size a can be a non-integer. Section 4 is a numerical evaluation of several sample designs using both fixed and random n_g.

2. Optimal Designs with Integer $\bar{n}_a$

The cost and variance models, (1) and (2), are algebraically of the same form as the cost and variance assumed in standard optimal allocation theory. Therefore, the values of m and n_a which minimise V for fixed $C = C_f$ are

$$\begin{aligned} m &= (C_f - C_0)\frac{\sqrt{RM^2C_1^{-1}}}{\sqrt{RM^2C_1}+\sum_{a=1}^{A}\sqrt{(1-R)N_a^2C_2}} \\ n_a &= (C_f - C_0)\frac{\sqrt{RM^2C_1^{-1}}}{\sqrt{RM^2C_1}+\sum_{a=1}^{A}\sqrt{(1-R)N_a^2C_2}} \end{aligned} \tag{3}$$

(Cochran, 1977, pp.96-99). The within-cluster sample sizes are therefore

$$\bar{n}_a = n_a/m_a = \sqrt{\frac{1-R}{R}\frac{C_1}{C_2}}. \tag{4}$$

If the intraclass correlations are low or the travel costs (C_1) are high, then the optimal $\bar{n}_a$ is high, so the sample is highly clustered. The number of clusters in sample depends on the cost constraint, C_f: if there are more funds available then a larger sample will be used. However, the within-cluster sample sizes, $\bar{n}_a$, do not depend on C_f, so this aspect of the sample design can be chosen without knowing the total budget available for the survey.

In practice, sample sizes must be whole numbers, but allocation (3) will generally give non-integer values of m, n_a and $\bar{n}_a$. The number of clusters, m, is usually large, so rounding m to the nearest whole number should work well. However, $\bar{n}_a$ is a different story, as it is a small integer, between 1 and a. Rounding of $\bar{n}_a$ may have a large effect. One impact of rounding is that the cost of the rounded design may be significantly different from the cost constraint C_f. Even if m is adjusted so that the cost constraint is met exactly, the resulting design is still not the best possible integer design. For example, suppose that all of the $\bar{n}_a$ in (4) are equal to an integer plus 0.4. Then all of the $\bar{n}_a$ would be rounded down, resulting in a much lower average sample size per cluster. It is possible that rounding some of the $\bar{n}_a$ up, and some down, will give a better solution.

To find the best integer-valued $\bar{n}_a$, notice that, for a given set of $\bar{n}_a$, m is determined by the cost constraint:

$$m = (C_f - C_0)\left(C_1 + \sum_{a=1}^{A} C_{2a}\frac{M_a}{M}\bar{n}_a\right)^{-1} \tag{5}$$

Substituting into equation (2) gives

$$\begin{aligned} V &= Rm^{-1} + (1-R)\sum_{a=1}^{A} n_a^{-1} \qquad (6) \\ &= Rm^{-1} + (1-R)\sum_{a=1}^{A}\left(\frac{M_a}{M}m\bar{n}_a\right)^{-1} \\ &= Rm^{-1} + (1-R)\sum_{a=1}^{A} MM_a^{-1}\bar{n}_a^{-1} \\ &= (C_f - C_0)^{-1}\left(C_1 + \sum_{a=1}^{A} C_2\frac{M_a}{M}\bar{n}_a\right) \\ &\quad \left(R + (1-R)\sum_{a=1}^{A} MM_a^{-1}\bar{n}_a^{-1}\right) \qquad (7) \end{aligned}$$

The integer optimal can be calculated by minimising (7) with respect to integer-valued $\bar{n}_a$. For each a, there are a possible values for $\bar{n}_a$. So there are $A!$ possible combinations of values of $\bar{n}_a$ where A is the maximum household size. In many cases, $A!$ is sufficiently small that (7) can be evaluated for every possible combination. For example, in the numerical study in Section 4, $A! = 6! = 720$. Notice that (7) does not depend on the cost constraint, so that the optimal integer sample size is independent of C_f, just like the optimal non-integer sample size.

3. Designs with Non-Integer $\bar{n}_a$

It is obviously not possible to select a non-integer sample size from a particular cluster. It *is* possible to allocate a range of integer sample sizes, n_g, to clusters g of the same size. Then $\bar{n}_a = n_a/m_a$ would be the average of these n_g, so that $\bar{n}_a$ can be a non-integer. It is proposed that n_g be randomly generated from an integer-valued distribution with $E[n_g] = \theta_a$.

A design of this type may give lower variances than the integer optimal design discussed in Section 2, because $\bar{n}_a$ can be set to be equal to, or closer to, the non-integer optimal values (4). However, the extra variation in n_g may result in greater variation in the estimation weights, which may increase the variance relative to the integer optimal design. It is unclear what the net result of these two factors will be: the numerical study in Section 4 compares the two approaches.

If n_g are random variables, there are at least two design unbiased methods of weighting the sample:

- The usual two-stage estimation weights for a cluster g of size a are $\dfrac{M}{m}\dfrac{a}{n_g}$.
- The inverse of the probability of selection for a unit i in cluster g of size a is

$$\begin{aligned}
\pi_i^{-1} &= P[i \in s] \\
&= \left\{P\left[g \in s^1\right] P\left[i \in s | g \in s^1\right]\right\}^{-1} \\
&= \frac{M}{m}\left\{E\left[P\left[i \in s | g \in s^1, n_g\right] | g \in s^1\right]\right\}^{-1} \\
&= \frac{M}{m}\left\{E\left[\frac{n_g}{a}\right]\right\}^{-1} = \frac{M}{m}\frac{a}{\theta_a}
\end{aligned}$$

The first weight uses the actual sample sizes, n_g, and the second weight uses the expected sample sizes θ_a. The first set of weights gives conditionally design unbiased estimators, the second set of weights gives a conditional bias, conditional on n_g. Both weights give design unbiased estimators, unconditionally over n_g. Clark (2002) derives weights which minimise the unconditional design variance; these turn out to be an interpolation between the two weights given above. It may seem more reasonable to use the first set of weights as they are conditionally unbiased, however this results in greater variation amongst the weights, which inflates the variance of estimators. As a result, the second set of weights often perform better.

The design variance for sample designs with random n_g will be denoted by V^*. The form of V^* is complicated and depends on whether inverse probability weights, weighting by $\dfrac{M}{m}\dfrac{a}{n_g}$, or an interpolation, is used. Clark (2002, ch.6) derives V^* and the optimal design-based weighting scheme. It is also shown that it is optimal to generate n_g as either a fixed integer θ_a with probability 1, or as a mixture of two neighboring integers. To illustrate, V^*, suppose that all clusters are of the same size, a, and that inverse probability weighting is used. Then

$$V^* \;=\; V + \phi^2 \frac{M^2}{m} a^2 R$$

where the relative variance of n_g is ϕ^2, for each g. See the Appendix for proof.

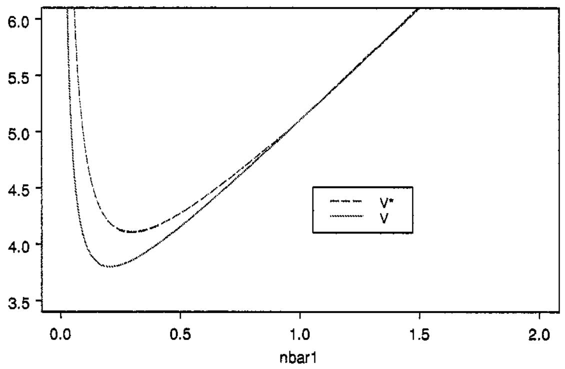
4. Numerical Study

A simpler way of calculating θ_a would be to approximate V^* by substituting θ_a for $\bar{n}_a$ in expression 7 for V:

$$\tilde{V} = \left(R + (1-R)\sum_{a=1}^{A} M M_a^{-1} \theta_a^{-1}\right).$$

Plot 1 illustrates the behaviour of V^* compared to $\tilde{V}$. The plot is based on a population consisting of an equal number of clusters of size 1 and size 2. One unit is selected from each selected cluster of size 2. The plot shows the behaviour of $\tilde{V}$ and V^* as the expected sample size from clusters of size 1, θ_1, is varied.

Plot 1: Variance for Fixed Cost with HHs of Size 1 and 2



The plot was calculated assuming: the optimal weighting method is used (as derived in Clark (2002)); $C_1/C_2 = 0.2$; $R = 0.4$; and the population variance for units in size 1 clusters is 0.25 times the population variance for units in size 2 clusters. The last assumption is not realistic and was made to exaggerate the difference between $\tilde{V}$ and V^*, for presentation. The optimal choice for $\bar{n}_1$ is about 0.3; if the approximation $\tilde{V}$ was used, $\bar{n}_1$ would be set at about 0.2. Notice that a value of θ_a less than 1 means that clusters of size a are subsampled.

Table 1 shows the variance of several alternative designs, for regression estimation of employment, with auxiliary variables agegroup by sex. Clusters are households and units are people. This table was calculated using data from the 1991 Australian Census of Population and Housing. Households contained between 1 and 6 adults. It is assumed that C_1/C_2 is about 0.25; this is probably somewhat lower than the ratio for face to face interviewing but may be appropriate for telephone interviewing. The intra-class correlation of employment, adjusted for age and sex, is about 0.2. All/household sampling is usually used in practice for labour force data. If the real-valued optimal θ_a, given by (4) in Section 2, are rounded to the nearest integer, the resulting variance is 0.96 (all variances are relative to the all/household design). The best integer optimal design, calculated using the method in Section 2, turned out to be half of the people in each household, rounding down; this gave a variance of 0.90. The best design allowing random n_g was found by numerically minimising V^* with respect to $(\theta_a : a = 1, \ldots, A)$ in Splus, using the NLMINB procedure (e.g. Venables & Ripley, 1994). This design took approximately half of the people in the household in expectation. Its variance was 0.87.

Table 1: Various Sample Designs, Regression Estimation of Employment

Design	Var.	θ_a					
		$a=1$	2	3	4	5	6
All/HH	1.00	1	2	3	4	5	6
One/HH	1.015	1	1	1	1	1	1
Rnded Real	0.955	1	1	1	1	2	2
Best Integer	0.896	1	1	2	2	3	3
Random n_g	0.867	0.5	1	1.5	1.9	2.1	2.7

Similar tables were calculated for different variables and different cost ratios. The best gains from the new methods are when C_1/C_2 is small, or when R is small.

5. Conclusions and Further Work

It is possible to make useful reductions in the variance in household surveys, by explicitly allowing for integer sample sizes. For the employment variable and a particular cost model, the variance was reduced by about 10% by using the best integer design rather than the usual all/household design. A further reduction of about 3% was made by allowing the expected within household sample sizes, θ_a, to be non-integers. In general, the new designs work best if the variable of interest is not highly correlated within households, and the costs associated with households are small compared to interview costs.

The new sample designs require interviewers to decide on the within-household sample size after identifying the size of the household. A randomly generated sample size may be required for some households. This complicates the interviewer's task but is probably feasible if computer assisted personal interviewing or telephone interviewing is used. Whether the gains justify the extra complication requires further study.

The main benefit of randomizing n_g seems to be to allow some subsampling of small households. It is counter-intuitive that an interviewer should knock on the door, find out the household contains only one or two people, and then terminate the interview with some probability. However, if the cost of this initial contact is small, it is sensible to devote resources to interviewing people from larger households, rather than have an excessive number of sole person households in sample.

This research could be extended to more sophisticated methods of sampling within households, for example stratification. Stratification within households has not been much used in practice, possibly because many strata would contain 0 or 1 units. The methods described in this paper could be used to design an effective within-household stratification scheme by allocating non-integer expected within-household sample sizes.

Appendix: Derivation of V^* for a Simple Case

The estimator is

$$\hat{T} = \frac{M}{m}\frac{a}{\theta}\sum_{g\in s^1}\sum_{i\in s_g} Y_i$$

where: Y_i is the variable for interest for unit i; s^1 is the sample of clusters g; and s_g is the sample of units i in cluster g. Let $\boldsymbol{n}$ be the vector containing all n_g for $g \in s^1$. Let S_g^2 and $\bar{Y}_g$ be the variance and mean, respectively, of Y_i over units i in cluster g. The design variance of $\hat{T}$ is

$$\begin{aligned} E_p\left[\hat{T}\right] &= E_p\left[var_p\left[\hat{T}|s^1,\boldsymbol{n}\right]\right] + var_p\left[E_p\left[\hat{T}|s^1,\boldsymbol{n}\right]\right] \\ &= E_p\left[\frac{M^2}{m^2}\frac{a^2}{\theta^2}\sum_{g\in s^1} n_g\left(1-\frac{n_g}{N_g}\right)S_g^2\right] \\ &\quad + var_p\left[\frac{M}{m}\frac{a}{\theta}\sum_{g\in s^1} n_g\bar{Y}_g\right] \end{aligned}$$

$$\begin{aligned}
&= E_p\left[E_p\left[\frac{M^2}{m^2}\frac{a^2}{\theta^2}\sum_{g\in s^1}\left(n_g-\frac{n_g^2}{N_g}\right)S_g^2|s^1\right]\right]\\
&\quad +E_p\left[var_p\left[\frac{M}{m}\frac{a}{\theta}\sum_{g\in s^1}n_g\bar{Y}_g|s^1\right]\right]\\
&\quad +var_p\left[E_p\left[\frac{M}{m}\frac{a}{\theta}\sum_{g\in s^1}n_g\bar{Y}_g|s^1\right]\right]\\
&= E_p\left[\frac{M^2}{m^2}\frac{a^2}{\theta^2}\sum_{g\in s^1}\left(\theta-\frac{\theta^2\left(1+\phi^2\right)}{a}\right)S_g^2\right]\\
&\quad +E_p\left[\frac{M^2}{m^2}\frac{a^2}{\theta^2}\sum_{g\in s^1}\theta^2\phi^2\bar{Y}_g^2\right]\\
&\quad +var_p\left[\frac{M}{m}\sum_{g\in s^1}a\bar{Y}_g\right]\\
&= \frac{M}{m}\frac{a^2}{\theta}\left(1-\frac{\theta\left(1+\phi^2\right)}{a}\right)\sum_{g\in s^1}S_g^2\\
&\quad +\frac{M^2}{m^2}\frac{a^2}{\theta^2}\theta^2\phi^2\sum_{g\in s^1}\bar{Y}_g^2+\frac{M^2}{m}a^2S_1^2
\end{aligned} \tag{6}$$

where S_1^2 is the population variance of $\bar{Y}_g$ over all clusters g. It is assumed that the mean of Y over all units, $\bar{Y}$, is zero, and that all clusters are of size a. In this case, the following identities hold:

$$\begin{aligned}
\sum_{g\in s^1}S_g^2 &\approx MS^2(1-R)\\
\sum_{g\in s^1}\bar{Y}_g^2 &= \sum_{g\in s^1}\left(\bar{Y}_g-\bar{Y}\right)^2\approx Ma^{-1}S^2(1+(a-1)R)\\
S_1^2 &= a^{-1}S^2(1+(a-1)R)
\end{aligned}$$

where S^2 is the overall population variance of Y_i and R is the population intraclass correlation of Y. The approximations are of order $1/M$. See for example Hansen et al. (1953). Substituting into (6) gives

$$\begin{aligned}
E_p\left[\hat{T}\right] &= \frac{M}{m}\frac{a^2}{\theta}\left(1-\frac{\theta\left(1+\phi^2\right)}{a}\right)MS^2(1-R)\\
&\quad +\frac{M}{m}\frac{a^2}{\theta^2}\theta^2\phi^2Ma^{-1}S^2(1+(a-1)R)\\
&\quad +\frac{M^2}{m}aS^2(1+(a-1)R)\\
&= \frac{M^2}{m}\frac{a^2}{\theta}S^2(1-R)-\frac{M^2}{m}aS^2(1-R)\\
&\quad -\phi^2\frac{M^2}{m}aS^2(1-R)+\phi^2\frac{M^2}{m}a^2aS^2(1+(a-1)R)\\
&\quad +\frac{M^2}{m}aS^2(1+(a-1)R)\\
&= \frac{M^2}{m}aS^2R+\frac{M^2}{m}\frac{a^2}{\theta}S^2(1-R)\\
&\quad +\phi^2\frac{M^2}{m}a^2aS^2R\\
&= V+\phi^2\frac{M^2}{m}a^2aS^2R.
\end{aligned}$$

References

Clark, R. G. (2002). *Sample design and estimation for household surveys* [PhD Thesis]. University of Wollongong.

Cochran, W. G. (1977). *Sampling techniques* (3 ed.). New York: Wiley.

Hansen, M., Hurwitz, W., & Madow, W. (1953). *Sample survey methods and theory vol.1 and 2.* New York: Wiley.

Venables, W., & Ripley, B. (1994). *Modern applied statistics with splus.* New York: Springer-Verlag.