OPTIMUM SAMPLE DESIGN FOR PERSONAL-VISIT ESTABLISHMENT SURVEYS

David W. Chapman, Klemm Analysis Group
1785 Massachusetts Ave., NW, 5th Floor, Washington, DC 20036

KEY WORDS: Optimization, Cluster Sampling, Stratified Sampling

1. Introduction

Large-scale, national surveys are generally classified into two types:

- Demographic or household surveys
- Economic or establishment surveys

Many national establishment surveys are conducted by mail or telephone. In such cases, the optimum sample design would not include clustering of establishments since there are no local travel costs associated with survey enumeration. For these types of surveys, optimum design generally focuses on Neyman allocation of the sample to strata, described by Cochran (1977, p. 99). Often the strata are defined by type and size characteristics.

However, for national personal-visit establishment surveys, serious consideration must be given in designing the sample to clustering the sample geographically, as is done with national household personal-visit surveys. A general discussion of the use of cluster sampling in the optimum design of personal-visit establishment surveys is given by Chapman (1993). Two examples of clustered national establishment surveys, the National Hospital Discharge Survey and the National Ambulatory Medical Care Survey, are described by McLemore and Bacon (1993).

For many establishment surveys, adequate national sampling frames (lists) are available commercially or from the survey sponsor. Therefore, it is often possible to define strata and allocate the sample to strata prior to sample selection, whether or not establishments are clustered geographically. In the case of a clustered establishment sample, the question arises as to whether or not Neyman allocation, applied before clusters are selected, still provides an optimum allocation of the sample to strata. The focus of this paper is the effect of clustering on optimum allocation of the sample to strata, and how these optima compare to Neyman allocation.

2. Basic Setup and Notation

Suppose that we have a national frame of N establishments (business locations), partitioned into L strata, the strata being defined perhaps by classifications of the establishments by size and type. In addition, the population of establishments is partitioned into M geographic clusters (e.g., defined by county) which cut across the stratum boundaries.

A two-stage probability sample of n of the N establishments is selected by first selecting a sample of m of the M clusters. The m clusters will be selected without replacement and either with equal or unequal probabilities. In the second stage, a stratified random sample of establishments will be selected from each of the clusters chosen at the first stage.

Letting "h" denote the stratum subscript, and "i" the cluster subscript, the following notation and basic definitions of terms will be used in subsequent sections:

N_{hi} = the total number of establishments in the sampling frame in stratum h in cluster i,

n_{hi} = the number of the N_{hi} establishments to be selected for the sample,

$\bar{x}_{hi}$ = the simple mean of a survey variable, X, computed for the n_{hi} sampled establishments,

N_h = the total number of establishments in the sampling frame in stratum h across all clusters,

$N_{h.}$ = the total number of establishments in the sampling frame in stratum h across the m sample clusters,

$n_{h.}$ = the total sample size for stratum h across the m sample clusters,

N_i = the total number of establishments in the sampling frame in cluster i across all strata,

n_i = the total sample size for cluster i across all strata,

S_{hi} = the population (unit) standard deviation among all establishments in the frame in stratum h in cluster i,

S_h = the population (unit) standard deviation among all establishments in the frame in stratum h across all clusters.

Consider an estimator, $\hat{X}$, of a population (frame) total. The optimization problem is to choose n_{hi} to minimize the variance of the estimator, subject to a fixed overall sample size n. In particular, we want to see what conditions are necessary so that the solution is equivalent to a Neyman allocation of the sample of n to the h strata, without regard to the sample of clusters selected.

3. Solutions to the Optimum Allocation Problem.

First, an attempt was made to derive an unconditional solution, that is, one which does not depend on the clusters selected at the first stage. Such a solution is best suited for comparison to the Neyman allocation sample stratum sample sizes, since those are unconditional. However, because of the complexity introduced by the dependence of each cluster allocation on the other clusters that are selected, an unconditional solution was not derived.

In terms of conditional solutions, two cases were analyzed: (1) m clusters selected with equal probability without replacement and (2) m clusters selected without replacement and with probability proportional to the total number of establishments in the cluster.

3.1 Clusters Selected with Equal Probability

With m clusters selected from all of the M clusters with equal probability and without replacement, an unbiased estimator, $\hat{X}$, of the population (frame) total for the variable X would be computed as follows:

$$\hat{X} = \frac{M}{m} \sum_{i=1}^{m} \sum_{h=1}^{L} N_{hi} \bar{x}_{hi} \qquad (1)$$

Applying the Lagrange Multiplier method, the optimum value of the stratum sample size, n_{hi}, within each selected cluster is:

$$n_{hi} = n \frac{N_{hi} S_{hi}}{\sum_{h=1}^{L} \sum_{i=1}^{m} N_{hi} S_{hi}} \qquad (2)$$

This solution is very similar to Neyman allocation, except that it is in terms of the stratum sample size within each selected cluster, rather than for the total stratum sample size. In order to derive a solution for the stratum sample size, $n_{h.}$, summed across all clusters selected for the sample, equation (2) was summed across the subscript i.

The resulting expression for the optimum value of $n_{h.}$ is still a function of within-cluster stratum parameters and does not resemble Neyman allocation. However, if it is assumed that the unit standard deviations across the m clusters for a given stratum, the S_{hi} values, are equal to the overall unit standard deviation for the stratum, S_h, the solution for the optimum stratum sample size, $n_{h.}$, summed across the m strata is no longer a function of within-cluster stratum parameters:

$$n_{h.} = n \frac{N_{h.} S_h}{\sum_{h=1}^{L} N_{h.} S_h} \qquad (3)$$

The optimum stratum sample size in equation (3) is very much like the Neyman allocation expression. The only difference is that it is based on the parameters for the m selected clusters, rather than on those for the entire population.

3.2 Clusters Selected with Unequal Probabilities

The second case analyzed is the selection of clusters with probability equal to the number of establishments it has in the frame. In this case, an unbiased estimator of the total for a variable X is:

$$\hat{X} = \sum_{i=1}^{m} \frac{N}{mN_i} \sum_{h=1}^{L} N_{hi}\, \bar{x}_{hi} \qquad (4)$$

To simplify the analysis, the estimator in equation (4) is based on the assumption that none of the clusters is large enough to be selected with certainty. The selection probability of any cluster is therefore equal to $m*(N_i/N)$, the relative size of the cluster times the number of clusters selected.

Again, applying the Lagrange Multiplier method, the optimum value of the stratum sample size within each cluster is:

$$n_{hi} = n \frac{(N_{hi}/N_i)\, S_{hi}}{\sum_{h=1}^{L} \sum_{i=1}^{m} (N_{hi}/N_i) S_{hi}} \qquad (5)$$

Comparing this solution to the one for the previous case given in equation (2), the term $N_{hi}S_{hi}$ is replaced by $(N_{hi}/N_i)S_{hi}$. These two equations would be equal only if the cluster sizes, N_i terms, were all equal.

To obtain an optimum stratum sample size for stratum h, summed across all of the m clusters, equation (5) was summed across i. As with the previous case, the resulting expression for $n_{h.}$ is still a function of within-cluster stratum parameters. In an attempt to eliminate these parameters, it was again assumed that the cluster standard deviations within a given stratum are equal to the overall stratum standard deviation. The resulting expression for the optimum stratum sample size, summed across the m sample clusters is:

$$n_{h.} = n \frac{S_h \sum_{i=1}^{m} (N_{hi}/N_i)}{\sum_{h=1}^{L} S_h \sum_{i=1}^{m} (N_{hi}/N_i)} \qquad (6)$$

The expression in equation (6) for the optimum stratum sample size is still a function of within-cluster stratum parameters. This makes it difficult to compare with the corresponding expression from the previous case, given in equation (3). If it assumed that the cluster sizes are equal, the expression in equation (6) reduces to that given in equation (3).

4. Conclusions and Recommendations for Future Research

The research reported here related to the optimum allocation of the sample to strata for geographically clustered establishment surveys. The optimum sample size solutions presented were conditional on the clusters selected at the first stage. Two cases were addressed: one in which clusters were selected with equal probability and the other for selecting clusters with probabilities proportional to the number of establishments in the cluster.

For the case of selecting clusters with equal probability, an expression for the optimum stratum sample size was derived that was very similar to Neyman allocation. However, the derivation required the assumption that the unit standard deviations were equal among all clusters within a given stratum. Also, the optimization expression, equation (3), was restricted to stratum sample sizes ($N_{h.}$ values) that were based only on the clusters selected.

For the case of selecting clusters with unequal probability, the optimum allocation expression, even with the equal standard deviation assumption (equation 6), did not resemble Neyman allocation . For the solution to resemble Neyman allocation, it is necessary to assume further that the cluster

sizes are equal.

The restriction of the derivations to optima that are conditional on the clusters chosen is a limitation of the results of this research since the sampling approach being considered was one in which the allocation to strata was made prior to selection of the clusters. It might be possible to derive optimum allocation solutions that are not conditional on the clusters selected. The derivation of such optima would probably require working with sampling fractions, rather than sample sizes, since sample sizes must be related to the set of clusters selected.

Even if unconditional optimum sample size solutions are derived, it seems unlikely that they would be independent of the within-cluster stratum parameters. At a minimum, assumptions about the equality of standard deviations across clusters within strata may always be required to obtain solutions that are not functions of within-cluster stratum standard deviations.

An obvious conclusion is that, in the case of geographic clustering for an establishment survey, it may not be prudent to apply an optimum allocation prior to selecting the clusters, even though the data to do so may be available. The optimum strategy may be to select the clusters first and then to consider the allocation to strata within the selected clusters.

This is not to say that stratum sample sizes should not be derived prior to sampling. In many cases, there are target precision requirements for national estimates of stratum parameters. Based on assumptions about design effects resulting from the geographic clustering and from the method of selecting clusters, an approximate stratum sample size needed to achieve the target precision level for a stratum estimate could be derived. Once the clusters are selected, the derived stratum sample size could then be allocated optimally to the clusters chosen.

An important area for future research relating to the optimum design of national personal-visit establishment surveys is the basic question of whether the sample should be geographically clustered. For national personal-visit household surveys, it is generally assumed that the optimum design involves geographic clustering at the first stage to avoid excessive local travel costs. However, for establishment surveys, it is not as obvious that geographic clustering is optimum because of the wide size variation of establishments and the more uneven geographic distribution of establishments. The answer to this basic question may depend on the specific type of establishment survey, the length of interview, the total sample size, cost parameters, and other factors.

REFERENCES

Chapman, David W. (1993), "Cluster Sampling for Personal-Visit Establishment Surveys." Proceedings of the International Conference on Establishment Surveys, June 27-30, 1993, Buffalo New York. American Statistical Association, Alexandria, VA.

Cochran, William G. (1977), Sampling Techniques, 3rd ed. John Wiley and Sons, New York, NY.

McLemore, Thomas and Bacon, Edward C. (1993), "Establishment Surveys of the National Center for Health Statistics." Proceedings of the International Conference on Establishment Surveys, June 27-30, 1993, Buffalo New York. American Statistical Association, Alexandria, VA.