Survey Designs to Optimize Efficiency for Multiple Objectives: Methods and Applications

Stephen Williams[1], Zhanyun Zhao[2], Frank Potter[2]
[1]Mathematica Policy Research, 148 Fletche Ct., Cary, NC, 27511
[2]Mathematica Policy Research

Abstract

Allocation of the sample among strata or sample clusters on the basis of variance components and survey costs is important to survey design. Optimum allocation equations for an estimated mean or total of a specific attribute and population have been known and used for years. However, we typically need an optimum allocation that simultaneously satisfies several types of estimates and for several variables and inference subpopulations. In this paper we review optimization methodology, its history and its extension to such multiple survey objectives. The computer algorithm we use to solve this nonlinear equation problem is described. Two recent applications are used to demonstrate the diversity of optimization problems and the flexibility and power of the methodology.

KEY WORDS: design optimization, multiple survey objectives, nonlinear programming

1. Introduction

This paper deals with the design of sample surveys. Stratification is a common feature of sample surveys, both to improve precision of survey results and to control sample sizes for important subpopulations. An important consideration in the use of stratification is how many sample units should be selected from each stratum with the objectives of either minimizing costs subject to precision requirements or maximizing precision subject to fixed resources. Also, we often use multi-stage designs, that is, selecting large units, then selecting smaller units within those first-stage units and so on, also referred to as hierarchal or cluster designs. The design task in this case is to determine the optimum number of units to select at each stage of sampling. These two techniques are often used together with stratification occurring at one or more of the stages. The equations for the optimal solutions to these problems when involving a single variable and parameter, such as average income, have been known and used for years.

However, we are usually interested in different types of estimates from sample surveys. A design that minimizes costs, for example, for one type of estimate often is not the best design, or even adequate for other estimates. The purpose of this paper is to present a method and examples of its use at Mathematica Policy Research (MPR) to achieve the required precision for multiple estimates at minimum cost. The applications are new but the method is not. Although the method is not new, it is not available for this specific application in most of the commercial statistical software presumably because of the difficulty of programming a global solution to all nonlinear programming problems.

2. Two Familiar Examples of Optimum Allocation

We first consider the familiar optimum allocation of the sample to strata that was mentioned above. The following solution by Neyman 1934 (see for example, Cochran 1964 p. 96) was obtained by a method in calculus known as Lagrange multipliers (Wider 1961 p. 136).

$$\frac{n_h}{n} = \frac{N_h S_h / \sqrt{C_h}}{\sum N_h S_h / \sqrt{C_h}}$$

This equation shows that the proportion of the sample, n_h/n, assigned to stratum h is larger for larger stratum populations, N_h, and larger stratum standard deviations, S_h, but smaller for larger stratum costs, C_h. By the same method, the following equation was obtained for the optimum allocation of the sample to the different stages in a multi-stage design.

$$m_{opt} \cong \sqrt{\frac{1-\rho}{\rho}} \sqrt{c_1 / c_2}$$

This reveals that the optimum cluster size, m_{opt}, in a two-stage design, such as selecting schools and then selecting students within the sample schools, should be larger with relatively small intra-cluster correlations, ρ, and if the cost of the first-stage units, c_1, is large relative to that of the units within the clusters, c_2 (this can be extended to more than the two-stages described here).

Again, these deal with a single type of estimator, attribute (or characteristic), and inference population. But as mentioned above, we are usually interested in multiple estimates requiring other methods of solving for optimum allocations. The more usual situation is

one in which we are interested in different types of estimates, such as, means, ratios, differences; and for multiple subgroups and characteristics in the same survey. The methods fall under the general heading of nonlinear programming (NLP).

3. Nonlinear Programming.

The field of NLP is complex and treated in many ways, essentially all methods involve some form of iteration process, like the Newton (or Newton-Raphson) method, and partial derivatives as with the Lagrange multiplier. The objective is to solve for the maximum or minimum of the objective function, F(x), often subject to constraints, either linear or non-linear. The constraints are of the form

$g_i(x) = 0 \quad (i = 1, ..., m1)$ where $m1 \geq 0$ and
$h_j(x) \geq 0 \quad (j = m1+1, ..., m)$ where $m \geq m1$

One of the greatest challenges in NLP is that some problems exhibit "local optima"; that is, spurious solutions that merely satisfy the requirements on the derivatives of the functions. One solution is to check when an optimum is indicated to make sure that a set of constraints are satisfied by this solution. Such a set of constraints was presented by Kuhn and Tucker 1951 and is used in the method discussed in this paper.

4. Nonlinear Programming Adapted to Hierarchy Sample Designs

When designing a sample survey, two important issues are quality of results and efficiency. In this application of NLP, we focus on efficiency. Two recent applications demonstrate the usefulness and flexibility of the software used at MPR. We recall one of the difficulties of developing a set of program steps for NLP is the diversity of applications. The paper by Chromy 1987 describes the basis for the software development that is now used at MPR. This software accommodates the classical sample designs involving hierarchical designs and stratification. Specifically, for stratification, the input involves a) the **cost function to be minimized**, which associates the cost components with the unknown sample sizes or a function of those sample sizes, x_h, for stratum h:

$$C = \sum_{h=1}^{H} C_h x_h + C_0$$

and b) the constraints including the **variance constraint** for each estimate of interest as a sum of variance components, V_{kh}, each divided by the sample size that influences that component, set equal to the maximum variance targeted for that estimate:

$$\sum_{h=1}^{H} V_{kh} / x_h + V_0 \leq V_k^*$$
$$and \quad x_h \geq 0$$

Note that both the cost equation to be minimized and the variance constraints have a term with subscript 0, which does not depend on sample size. These terms are excluded from the problem.

The function to be minimized can be expressed as:

$$F(\underline{x}) = \sum_{h=1}^{H} C_h x_h + \sum_{k=1}^{K} \lambda_{ik} \sum_{h=1}^{H} V_{hk} / x_h$$

where C and V are the cost and variance components,
$\underline{x}$ is the vector of sample size functions, and
λ is the Lagrange multiplier.

The Kuhn-Tucker conditions that must be satisfied for this to be the minimum cost solution subject to constraints are:

$\partial F(\underline{x}) / \partial (x_h) = 0$ for h=1 to H partial derivitives;
$\lambda_{ik} \geq 0$ for iteration i; k=1 to K;
$x_{ih} \geq 0$ for iteration i; h=1 to H;
$\lambda_{ik} \; [V_k^* - V_{ik}] = 0$ for iteration i and all k, and
all variance constraints, V_k^*, are satisfied.

5. An Example Application

In this example application, the fifth round of a large national survey of physicians is being redesigned from a telephone-only survey to a mixed mail and telephone data collection mode. The existing design uses a stratified multi-stage probability sample. Specifically, the setting for this application is the redesign of the fifth round of the Community Tracking Study (CTS) Physician Survey. CTS is a national study of the rapidly changing health care market and the effects of these changes on physician practices. Funded by the Robert Wood Johnson Foundation, the study is conducted by the Center for Studying Health System Change (HSC). (Information about other aspects of the CTS and HSC is available at www.hschange.com.) When changing methods in a continuing type of study such as this we need to be concerned about the interruption in the trend analyses that results from the method change.

For the proposed redesign for round five, a mode transition survey, there are three basic strata in the telephone component (the two modes are used concurrently to assess the impact of the mode change on the trend analyses). One of the strata used a clustered sample with unequal sampling rates that produces an equal probability sample of physicians. The second stage selection probabilities had to reflect the fact that primary sampling units were selected with probability proportional to 1990 Census of Population counts because the same PSUs are used. The other two strata use un-clustered stratified sampling and the mail survey component also uses an equal probability stratified sample.

We were interested in controlling precision for seven different estimation equations, including ratio estimators, difference estimators, and direct expansion estimators for each data collection mode and combined. Also, there were 21 key variables of interest identified for these estimators. Some of the variables were discrete and some continuous.

5.1 Formulation of the Problem

The problem must be formulated to be consistent with the equations presented in Section 4, above. The objective function is the cost function that is to be minimized. For the variance constraints, an average of the population variances for the 21 key variables was used to reduce the number of variance constraints to 7 (one for each estimator form). The values for the variance components and other parameters of the problem were based on information from previous rounds of the survey. Once the optimum solutions were obtained based on the 21 variable averages, some of the more important variables were checked individually to make sure the allocation satisfied their precision constraints. The 7 estimates related only to the total population (no subpopulations) and included a mode difference estimate, a ratio change estimate, and 5 means (modes combined without adjustment, combined and adjusted to mail mode responses, combined and adjusted to telephone mode responses, using only mail, and using only telephone). The cost function contained 6 terms (3 strata times 2 modes) reflecting the estimate based on a prior experiment that indicated a telephone response costs three times a mail response in this setting.

The object equation is:

$$C = n_{11} + n_{12} + n_{13} + 3n_{21} + 3n_{22} + 3n_{23}$$

The sample sizes being sought, nij comprise the x vector terms noted in the general formulation of NLP problems. The six terms also coincide with the six terms in each of the seven variance constraint equations

The first of seven estimators is used to demonstrate the development of the seven variance equation constraints.

$$\hat{\bar{Y}}_{Mf} = \sum_{h=1}^{3} W_h \lambda \bar{y}_{1h} + W_h (1-\lambda) \bar{y}_{2h} \hat{R}$$

$\bar{y}_{ih}$ = estimated mean for stratum h based on i^{th} mode (1 = mail, 2 = telephone),
W_h = weight for stratum h,
λ = a weight ($0 \le \lambda \le 1$), based on relative costs, for combining the two mode means, and
$\hat{R}$ = estimated ratio of reported means, mail over telephone mode.

This is the estimated mean based on the combined samples with the telephone responses adjusted to the mail mode level. .

The first of six components for the variance of this estimator is (the one divided by n11, the number to be assigned to mail mode in stratum 1):

$$VC_{11} = W_1^2 [Deff_1^2 S_{11}^2 + 2\rho_{MT} \lambda \sqrt{Deff_1} S_{11} (1-\lambda) \sqrt{\frac{1-\lambda}{\lambda} Deff_1} S_{21}^2 (\hat{R}^2 - V_R^*)] / n_{11}$$

Where $Deff_1$ = average design effect for stratum 1,
S_{ij} = average relative standard deviation for mode i and stratum j,
ρ_{MT} = correlation between mode means within a stratum,
V_R^* = relative variance of $\hat{R}$, and

other terms as defined above.

Table 2. Sample Allocation Based On Initial Solution

Stratum	Sample Counts
1-mail	2,369
1-tele	1,368
2-mail	942
2-tele	544
3-mail	770
3-tele	445
Total	6,437

As noted before, there are seven variance constraint equations, all different, and each with six components such as the one just shown (6x7=42 equations).

5.2 Results of the Optimization

The input values for the 6 variance components, for the first of 7 estimators are presented as examples in Table 1. The target variance is 0.0023 (the Vo subtracted from the overall variance constraint is that part of the variance that is not influenced by sample size).

The sample allocation based on the initial variance constraints is shown in Table 2. The sample sizes for stratum 1 are largest reflecting the fact that approximately 60 percent of the population is represented by that stratum. The larger sample to the mail sample compared to the telephone in each stratum reflects the anticipated cost differences.

Table 3. Target and Resulting Variances of the Seven Estimates for the Initial Solution

Estimate	Target	Result Var
Mail, combined	0.0023	0.0007
Tele, combined	0.0022	0.0007
Unadjusted combined	0.0025	0.0007
Mode difference	0.0025	0.0025
Change	0.0014	0.0007
Mail only	0.0025	0.0010
Telephone only	0.0025	0.0017

The total sample size of 6,437 required to meet the constraints in the first solution attempt was slightly larger than desired. Therefore, the next step was to compare the target and resulting variances for each of the estimates (Table 3).

Since the intent was to obtain a slightly smaller total sample, we identify the estimates(s) that have equal target and resulting variance values. We must accept a slightly lower precision for these estimates in order to reduce the total sample size. Hence, we reduce the variance constraint for those estimates and rerun the optimization program.

We note in this case that the resulting variances all surpass their target variance (slack constraints) except for the mode difference estimator. In order to reduce the total sample size required by the first solution, therefore, we need to relax the variance constraint for that estimate.

Table 1. Variance Component Values for the Full Sample Adjusted Mail Mode

Estimate	Variance Component
Y_{Mf}	0.518 /n11
	0.222 /n21
	0.079 /n12
	0.029 /n22
	0.053 /n13
	0.020/ n23
Target	0.0025-Vo=0.0023

6. Application Two

At MPR , we have used this optimization method in the design of numerous surveys, but a brief discussion is presented for one other application for the purpose of demonstrating the flexibility of the method. In this second application, a different focus on optimization for allocation to over 125 demographic/program-participant categories for which specific precision requirements were to be met.

In terms of the sampling design, the survey included two stages of selection and used a composite size measure for selection of the first stage units, the primary sampling units (PSUs) (Folsom et al. 1987). A PSU was a single county or a group of adjacent counties. The frame contained more than 500 PSUs (some 300 MSAs and 200 PSUs formed from non-metropolitan counties).

The composite size measure was computed for each PSU using the count of program participants in the PSU in each of eight sampling strata. The size

measures were designed to permit constant global selection rates of participants for each of the eight sampling strata and to equalize survey workload in each first stage unit. The PSUs were then selected with probabilities proportional to this composite size measure.

The process of developing an appropriate allocation across the sampling strata of participants was based on the evaluation of alternative sets of constraints for estimates for the population domains. More than 125 precision constraints were developed for 9 to 11 analytic domains. The constraints were developed by reviewing the analytic importance for combinations of analysis domains and subpopulations (for example, gender). Through an iterative process of alternative sets of constraints, a series of acceptable constraints for the analysis domains and subpopulations were developed.

In preparing the revised sample allocation, additional constraints were added to ensure precision for estimates for populations defined by age and living situation of program participants (the initial constraints only included current age and living situation). Constraints were imposed and evaluated based on the need for precision for specific analytic domains.

The development of the allocation of the sample among the sampling strata uses statistical models for the sampling variance for the specific estimates for various domains and precision constraints for these estimates. The derivation of the variance model for this survey design was:

$$\text{Var}\,(p \mid d, h, s) =$$

$$\Sigma_h \, W_{hd}^{\,2} \, [1 + \rho \, (e_d \, s-1)] \, p \, (1-p) \, / e_d \, n_h$$

where

- d denotes the analytic domain
- h denotes the sampling strata based on participant classifications
- n_h denotes the respondent sample size in stratum h
- s is the number of respondents in each PSU
- ρ is the intra-cluster correlation among responses to a dichotomous variable (ρ is projected at 0.01)
- e_d is the estimated proportion of respondents in a stratum that are in analytic domain d.

The intra-cluster correlation was projected at 0.01 because the diversity of participant characteristics are likely to make the responses be less homogeneous within a PSU than in some household surveys of a similar population.

The variance model incorporates stratification, the eligibility or membership in a domain of interest for specific respondents, and the clustering of respondents within PSUs.

Various algorithms have been used to minimize a cost function (the objective function) with respect to multiple precision constraints (a system of equations) in the mold of linear programming. One such method is described above as used in the first application and was also used for this study.

7. Summary

For smaller surveys or surveys with relatively few statistics of importance to the results, methods such as those described in Section 2 using familiar optimization equations are often all that is needed for the design. On the other hand, these basic equations are often inadequate for designing large complex surveys that must ensure specified levels of precision for a range of questions, different forms of estimates, or for numerous inference subpopulations, In this paper we describe a very flexible and powerful method for dealing with multiple objectives in complex sample surveys using simultaneous solutions. Two recent applications were presented.

The specific software for using this method is not available in many of the sampling software packages that are commercially available, but MPR and other researchers have programmed the algorithm in SAS or by providing the appropriate equations as input to the SAS Proc NLP. The paper by Chromy (1987) provides sufficient description of the steps to facilitate programming the method. Also SAS NLP can be used for some applications by restructuring the specific problem into a function of functions of the objective variables (number of functions must be equal to number of objective variables for the problem to have a consistent and unique solution). Both linear and nonlinear, equality and inequality constraints can be entered as well as a choice of the method for solving, for example, Newton-Raphson method with ridging is the default if *TECH=* is not specified. The number of iterations required for convergence increases as linear constraints are added and again when both linear and nonlinear constraints are added.

As can be seen in the applications presented here, the tedious feature of application can be calculating the variance components for the variance constraint equations. Once the problem is programmed and variances developed, however, additional options can be quickly explored for a given setting.

References

Chromy, James R.(1987). "Design Optimization with Multiple Objectives" In *Proceedings of the American Statistical Association,* Alexandria, VA.

Cochran, William G. (1964), *Sampling Techniques*, Wiley and Sons, New York.

Folsom, R.E., F.J. Potter, and S.R. Williams (1987) "Notes on a Composite Size Measure for Self-Weighting Samples in Multiple Domains." In *Proceedings of the American Statistical Association,* Alexandria, VA: American Statistical Association.

Kuhn, H.W. and A.W. Tucker (1951) "Nonlinear Programming". *second Berkley Symposium on Math Statistics and Probability*, J. Neyman, ed., University of California Press, Berkeley, California, 481-492.

Wider, D.V. (1961), *Advanced Calculus*, Prentice-Hall, Englewood Cliffs, NJ.