OPTIMUM CLUSTER SIZES IN MULTI-PURPOSE SAMPLE-SURVEYS

Babubhai V. Shah and Ralph E. Folsom, Research Triangle Institute

I. INTRODUCTION

The NAEP (National Assessment of Educational Progress) in school sampling design is a six-stage design. The six stages are: (1) Region, (2) Size of Community (SOC), (3) Cycle, (4), PSU, (5) School, and (6) Student. The third stage, cycle, represents a set of pseudo-strata formed by collapsing state substrata nested within major Region x SOC categories. These psuedo-strata (cycles) were introduced to facilitate the calculation of standard errors for NAEP statistics. The first three stages are assumed fixed stratification levels and are, therefore, not subject to change. Thus, the problem of finding the optimal design is reduced to finding the configuration of PSUs, schools, and students that will provide minimum variance (maximum efficiency) at a given cost. Since the number of PSUs, schools and students are constrained by the total cost, the two independent parameters of the NAEP design are (1) the number of schools per PSU and (2) the number of students per school. The basic objective of the study has been to determine the "optimal" values of these two parameters.

To determine the optimal design, estimation of variance and cost components was required. A detailed study of the cost components for NAEP's Year-02 design (Working Paper No. 8) was available and it was decided to use the results of this study since the Year-02 design and data collection procedures closely parallel the Year-03 and 04 assessments. The relevant details are presented in Reference [3]. For the estimation of the variance components, two computer programs were adapted for the NAEP design and compared. One, using the formula by Henderson [4], which is available through the Statistical Analysis System [1] was compared to "VARCOMP", an RTI algorithm developed by Shah [7] using a formula by Seeger [6]. Computation of the variance components for several statistics indicated little numerical difference between the two techniques. However, the cost of computing variance components through SAS was approximately 25 to 50 percent higher than that by "VARCOMP". The details of the formula used in VARCOMP appear in the working paper [7].

With respect to optimality criteria, if one is interested in estimating only one statistic then the solution for the optimal design is well known [5]. However, no well-defined solution exists for the "optimal" design for many statistics. A feasible definition is developed in Section II. The results for NAEP designs for group packages are presented in Sections III and IV.

II. OPTIMALITY CRITERION

If the objective of a sample survey is to estimate only one statistic, then the usual optimality criterion is the minimum variance for the statistic at a given cost. However, it is rare, in any survey, that one is interested in only one statistic. The optimality criterion for many statistics is not quite obvious. Some possible suggestions are (a) the design that is optimal for most statistics, (b) the design that has minimum average variance at the given cost, and (c) the design with maximum average efficiency.

The average of several quantities is meaningful only if all the quantities are measured on the same scale and units. The variances of different statistics would be measured on different scales and units and, hence, the minimum average variance does not appear to be a meaningful criterion. To avoid the problem of proper scale, it is appropriate to define the efficiency of a design for a statistic. The efficiency of a design is a pure ratio with the numerator equal to the minimum variance that can be achieved by the optimal design for that statistic and the denominator is the variance of the same statistic for the given design.

The objective is to find the design with maximum average efficiency at the given cost and it would be desirable to have as small a variance of efficiencies over all statistics as possible. The trade-off between the maximum mean and minimum variance of efficiencies is not easy to define. However, in practice if the optimum is stable, then we may regard the minimum variance as a secondary criterion for selecting from several designs which are nearly optimal. It should be noted here that an ideal theoretic approach would be to consider the appropriate multi-variate distribution of many statistics.

It is not possible to obtain an explicit solution for the design with maximum average efficiency. Hence, an indirect attempt to solve the problem will be made. Moreover, the practical limitation on the sample survey design is likely to reduce the number of feasible designs to a few; for the way the cost model is defined and derived, it will be appropriate for only a few designs in the neighborhood of the current design. From a practical point of view, it will be sufficient to evaluate means and variances of efficiencies over all statistics for these few feasible designs, in order to determine the "optimal" design from among the practically feasible designs.

Let us assume there are M designs D_i, $(i = 1, 2, \ldots, M)$ and N statistics Y_j, $(j = 1, 2, \ldots, N)$. Let the estimates of the variance components of Y_j for PSU, school, and student be denoted by $V(j, \ell)$, $(\ell = 1, 2, 3)$ respectively. The details regarding definitions and procedures for deriving variance components are given in Reference 8. If the cost function is assumed to be linear and the variable unit costs for PSU, school, and student are c_1, c_2, and c_3 respectively, then the efficiency $E(i, j)$ of the ith design which has p_i PSUs, s_i schools per PSU, and k_i students per school can be derived to be

$$E(i, j) = \frac{\text{Minimum Variance at the given cost}}{\text{Variance for the given design}},$$

where

$$\text{Numerator} = \{\sum_{\ell=1}^{3} \sqrt{c_\ell V(j,\ell)}\}^2 \div (c_1 p_i + c_2 p_i s_i + c_3 p_i s_i k_i)$$

and

$$\text{Denominator} = \frac{V(j,1)}{p_i} + \frac{V(j,2)}{p_i s_i} + \frac{V(j,3)}{p_i s_i k_i} .$$

The expression for minimum variance at given cost is given in any sampling textbook, e.g., Murthy [5]; the expression for the variance of a given design is developed in Reference [8]. Once the efficiencies E(i, j) are computed, means and variances over j can be computed as

$$M_i \{E(i,j)\} = \frac{1}{N} \sum_{j=1}^{N} E(i,j),$$

$$V_i \{E(i,j)\} = \frac{1}{N-1} \sum_{j=1}^{N} \{E(i,j) - M_i\}^2 .$$

These means (M_i) and variances (V_i) have been used in this report for determining the optimal design.

III. RESULTS

The data used for this study consists of six packages: two group packages for each of the three age groups. For each appropriate1/ item of these packages, forty-one (41) sets of variance components were estimated: one set associated with national p-values (the national estimated proportion of correct or acceptable answers), twenty (20) sets associated with p-values for twenty (20) domains, and twenty (20) associated Δp-values (Δp-value equals domain p-value minus the national p-value). The twenty domains that were considered are as follows: 2 by sex, 2 by race, 5 by parents' education, 4 by region, and 7 by STOC: (Size and Type of Community).

An extensive study has been carried out by Folsom and Hartwell [3] to specify a cost model for a general NAEP In-School Survey. The cost estimates were used to define 15 designs with the same costs. These designs for group packages are given, and their efficiencies are presented in Table 3.1. No efficiencies were computed for those statistics with domain p-value equal to 1 or 0. The results in Table 3.1 are based on 7578 efficiencies.

IV. COST SENSITIVITY ANALYSIS

Doubts have been raised about the accuracy of the cost estimates. It was felt that some study of possible variation in the efficiencies and "optimal" design parameters resulting from varying cost components was necessary. It should be noted here that it is not necessary to investigate the effects of such inaccuracies arising from sampling variation in the estimates of our variance components, since we are averaging over several thousand estimates. The standard deviations among estimated efficiencies are presented in Table 5.8. If the statistics were independent the standard error of a mean efficiency would be smaller than .05 percent.

Various knowledgeable persons at RTI have indicated that the errors in allocation of variable costs to PSU, school, and student are not likely to be more than 20 percent. The critical parameters that effect the optimality of designs are ratios of the variable cost components, such as (variable cost per PSU) ÷ (variable cost per school). If one of the costs is increased by 20 percent and the other is reduced by 20 percent the ratio will be altered by 1.5 or 0.66. To study the effect of the possible errors in the cost, the mean efficiencies were recomputed for all 15 designs. In each design, the number of schools per PSU and number of students per school were kept the same; however, the total number of PSUs was adjusted to make it consistent with the change in cost-ratios.

The results indicate that the mean efficiencies do not vary more than five percent under cost fluctuations within the above range. The previously obtained optimum design remains optimal or is second or third best and is within 0.5 percent in efficiency compared to the "best" one for the cost ratio. Thus, we feel that the optimal design is quite stable under cost changes.

REFERENCES

[1] Barr, Anthony James and James Howard Goodknight. A User's Guide to Statistical Analysis System. North Carolina State University, Raleigh, North Carolina 27607, 1972.

[2] Folsom, Ralph E., David L. Bayless, and B. V. Shah. Jack-knifing for Variance Components in Complex Sample Survey Designs. Presented at the American Statistical Association Meetings at Fort Collins, Colorado, August 23, 1971.

[3] Folsom, Ralph E. and T. D. Hartwell. A Study of Optimum Designs for an In-School Assessment, Working Paper No. 8, RTI Project 25U-688, 1972.

[4] Henderson, C. R. Estimation of Variance and Covariance Components, Biometrics, Vol. 9; 1953, pp. 226-252.

[5] Murthy, M. N. Sampling Theory Methods. Statistical Publishing Society: 20411 Barrackpore Trunk Road, Calcutta-35, India, 1967, p. 347.

[6] Seeger, Paul. A Method of Estimating Variance Components in Unbalanced Designs. Technometrics 12, 1970, pp. 207-218.

[7] Shah, B. V. and C. N. Dillard. VARCOMP Program for Computing Variance Components. Working Paper No. 30, RTI Project 255U-796, 1974.

[8] Shah, B. V.; Folsom, R. E. and Clayton, C. A. Efficiency Study of YEAR-03 In-School Design. RTI Final Report 255U-796-2, 1973.

Table 3.1

Mean and Standard Error of Efficiencies for 15 Cost Equivalent Designs for NAEP Study

Design	No. of PSUs	No. of Schools per PSU	No. of Students per School	Total No. of Students	Total Variable Cost	All	Standard Deviation
1	157	1	16	2,512	303764	69.9	13.8
2	179	1	12	2,148	303004	66.5	14.5
3	209	1	8	1,672	303200	59.1	15.4
4	85	2	18	3,060	300475	77.0	11.3
5	92	2	16	2,944	302952	77.2	11.6
6	99	2	14	2,772	302041	77.0	12.1
7	108	2	12	2,592	303359	76.1	13.0
8	118	2	10	2,360	302887	74.2	14.2
9	130	2	8	2,080	302224	70.7	15.6
10	60	3	18	3,240	300851	77.8	11.1
11	65	3	16	3,120	302323	78.4	11.1
12	70	3	14	2,940	300164	78.7	11.5
13	77	3	12	2,772	302225	78.5	12.3
14	85	3	10	2,550	302765	77.2	13.5
15	95	3	8	2,280	303894	74.5	15.1