OPTIMAL ALLOCATION FOR MULTIPLE-OBJECTIVE SAMPLES

Eugene C. Poggio, Abt Associates Inc.

ABSTRACT

Many samples are used to estimate more than a single quantity. A sample might be used to estimate several population means (or totals), means for several subpopulations or treatment groups, differences between means for various subpopulations or treatment groups, means for both primary and secondary stage units, or, frequently, combinations of these. The sample allocation for such multiple-objective samples is often based on only a single quantity, a mean or proportion for the total population. This paper proposes a simple basis for allocation of samples that are to be used for multiple objectives--a weighted sum of the standard deviations or of the coefficients of variation of the estimators of the quantities of interest, where the weights are chosen to reflect the relative importance of the various quantities to be estimated. The optimal allocation is determined by minimization of the weighted sum subject to a cost constraint. Practical procedures for performing the minimization are discussed, and several applications are examined.

1. INTRODUCTION

Many samples, if not most, are used to estimate more than a single quantity. A sample might be used to estimate several population means (or totals), means for several subpopulations or treatment groups, differences between means for various subpopulations or treatment groups, means for both primary and secondary stage units, or, frequently, combinations of these. Further, the value of the precision of estimates often differs substantially among the various quantities to be estimated, as for example between an overall mean and a subpopulation mean.

The sample allocation used for such a multiple-objective sample is often in practice simply the optimal allocation for some single selected "primary" objective. Although the optimum is frequently relatively flat so that little efficiency is lost by allocations reasonably close to the optimum allocation (Cochran (1963)), there are instances in which losses in efficiency resulting from such an approach may be substantial.

Several authors have suggested approaches to address the problem. Bean and Burmeister (1978) provide an excellent review of a number of approaches proposed for allocating a stratified sample to estimate several population means or totals. Although much of the literature is confined to this context, most of the approaches (if not the solutions) are applicable to the broader context presented above. In reviewing the literature in this paper, all of the proposed approaches will be presented in this wider context.

Cochran (1963) considered the approach of calculating the optimal allocations for a specified sample size for each of the quantities to be estimated and using the averages across these allocations as a compromise allocation. In an example he provides in which three population means are estimated with a stratified sample, the approach performs quite well. There is, however, no theoretical basis for the approach and there are undoubtedly cases, especially in the broader context considered herein, in which it does not perform so well.

Yates (1953) proposed allocating the sample to minimize the cost subject to the constraint that the variances of the estimators be equal to specified values. Dalenius (1953, 1957) relaxed the unnecessarily restrictive constraint by allowing the variances of the estimators to be less than or equal to the specified values. A disadvantage of their approach is the difficulty of specifying maximum values for the variances of each of the estimators. Also, the approach may result in poor tradeoffs; some maxima may be very costly to achieve and the data collector may well prefer to use scarce resources to make greater reductions in other variances.

A variant of the above approach which attempts to provide a practical solution to the problem of setting the variance tolerances and to consider tradeoffs was suggested by Chatterjee (1968). This approach also begins by specifying a set of variance tolerances and minimizing the cost subject to the constraint that the variances of the estimators do not exceed the tolerances. The "shadow price" for each of the estimators (that is, the partial derivative of cost with respect to the specified tolerance for that estimator) is then calculated giving the expected decrease in cost for a small decrease in the variance tolerance. The shadow prices are then used to calculate the cost of a number of sampling plans. This provides the data collector with a series of plans with varying cost and precision from which to choose. This approach has the distinct advantage of taking into consideration the various tradeoffs between cost and precision. However, the procedure is without any formal basis, and it may still be difficult to specify the set of variance tolerances at the outset.

A different approach proposed by Chakravarti (1955) and Dalenius (1957) and examined further by Ghosh (1958) is to minimize the generalized variance (the determinant of the variance-covariance matrix) subject to a cost or sample size constraint. Aoyama (1963) approached the problem from the perspective of minimizing the area of the ellipsoid of concentration, which reduces to minimizing the generalized variance.

This approach has the advantage of not requiring specification of variance tolerances. On the other hand, it does not allow one to take into account any differences in the values of precision for the variances and covariances of the estimators. Further, it may not be appropriate to take into consideration at all the values of the covariances.

Several authors have examined approaches based on the ratios of the actual variances of the estimators to the minimal variances under optimal allocations for each estimator. These ratios provide a measure of the relative loss of efficiency resulting from the compromise allocation. Dalenius (1957) proposed, and Chatterjee (1967) and Kish (1976) examined further, the approach of minimizing a (weighted) sum of these ratios subject to a cost or sample size constraint. Peters and Bucher (undated) instead proposed maximizing the sum of the inverses of these ratios. These approaches avoid the need to specify variance tolerances and do address tradeoff issues. However, because the quantity minimized (or maximized) in these approaches may differ substantially from the data collector's actual objective function, the resulting allocation may not be optimal in terms of the data collector's objectives.

For data collection in which errors of estimates can be considered losses measured in monetary units, Yates (1960) proposed allocating the sample so as to minimize the sum of the expected total loss and the data collection costs. He considered specifically the case in which the expected loss can be expressed as a linear function of the variances of the estimated population means or totals. Cochran (1963) also discusses this approach. This approach is elegant and well-suited to this problem. Typically, however, one cannot define losses explicitly.

Suggesting a basis for allocation related to Yates' approach, Hartley (1965) considers minimization of a weighted sum of the variances of the quantities to be estimated subject to a cost constraint. This approach differs from Yates' both in its treatment of data collection cost as a constraint, rather than as a term in the quantity to be minimized, and by not explicitly interpreting the weighted sum as an expected loss. The approach proposed in this paper is similar to Hartley's approach.

2. PROPOSED APPROACH

As a basis for allocation of a sample that is to be used to satisfy several objectives, it is here proposed that a weighted sum of either the standard deviations or the coefficients of variation of the estimators of the quantitites of interest be minimized subject to a cost constraint. The weights are chosen to reflect the relative value of precision for the various quantities.

Specifically, suppose one desires to estimate a set of quantities q_i, using estimates $\hat{q}_i$ with variances σ_i^2. Let w_i denote the weights corresponding to q_i. Let $f(\underset{\sim}{c}, \underset{\sim}{n})$ denote a data collection cost function which reflects all of the cost components ($\underset{\sim}{c}$) and sample size components ($\underset{\sim}{n}$) of the sample design (i.e., c_h = cost per unit in stratum h, and n_h = number of units in stratum h), and let C denote the maximum total cost for data collection. The proposed allocation procedure is then to find the $\underset{\sim}{n}$ which minimizes either the sum

$$S = \sum_i w_i \sigma_i \quad \text{or the sum} \quad S' = \sum_i w_i \frac{\sigma_i}{q_i} ,$$

subject to the cost constraint $f(\underset{\sim}{c}, \underset{\sim}{n}) \leq C$ and to population size constraints $n_h \leq N_h$. When the cost components are all equal, the cost constraint reduces to a sample size constraint of the form $N = \sum_h n_h$.

Use of the standard deviations of the estimators is, for large sample sizes and unbiased estimators, approximately equivalent to use of the expected confidence interval length, assuming (1) asymptotic normality and (2) equal confidence coefficients for all confidence intervals. Similarly, use of the coefficients of variation of the estimators is approximately equivalent to use of the relative length of the expected confidence interval measured with respect to the quantity being estimated. The latter approach is particularly useful when several items measured in different units are being estimated since the coefficient of variation is unit-free. If the former approach is used when items measured in different units are being estimated, the weights must take into account the units in which each quantity is measured. If estimators are biased, mean square errors should be substituted for the standard deviations.

The procedure is general in that, at least in principle, it provides a formal basis for optimally allocating any type of sample to be used to estimate any set of quantities. It can be used to allocate stratified or multistage samples from which population means, proportions, or totals for several items are to be estimated. For a multi-stage sample, the quantities to be estimated might include means for primary, as well as secondary, stage units (e.g., for hospitals, as well as for patient medical records). It can be used to allocate samples to be used to estimate means for several subpopulations or treatment groups, differences between these means, or both.

The solution to the minimization can be accomplished in several ways. Analytic solutions are often obtainable in cases that are comparatively simple. Often symmetries in the problem can be used to simplify the solution. In some cases, direct enumeration of cases on a computer provides a feasible and simple solution. For each possible case allowed by a cost constraint, the value of S (or S') can be computed and the allocation yielding the smallest value of S selected as the optimum. Usually computational costs can be reduced substantially by eliminating large classes of cases which can be determined a priori not to be optimal. Another device that can be used to reduce costs is to consider only cases in which component sample sizes are multiples of some small integer, thus providing an approximate solution to the optimization. Finally and in general, a solution can be obtained using nonlinear programming methods.

3. EXAMPLE 1

Suppose one is designing a sample in which there are three treatment groups and one control group and that one is equally interested in the three differences in means between each treatment group and the control group:

$$\mu_{T1} - \mu_C \qquad \mu_{T2} - \mu_C \qquad \mu_{T3} - \mu_C$$

Assuming that (1) the differences are estimated as the differences between sample means and (2) the costs per unit of observation are equal among the four groups, the proposed procedure would minimize the sum

$$S = \left(\frac{\sigma^2_{T1}}{n_{T1}} + \frac{\sigma^2_C}{n_C}\right)^{1/2} + \left(\frac{\sigma^2_{T2}}{n_{T2}} + \frac{\sigma^2_C}{n_C}\right)^{1/2} + \left(\frac{\sigma^2_{T3}}{n_{T3}} + \frac{\sigma^2_C}{n_C}\right)^{1/2},$$

subject to the sample size constraint $n_{T1} + n_{T2} + n_{T3} + n_C = N$, where the σ^2's and the n's denote the variances and sample sizes corresponding to each treatment and comparison group.

Lacking prior information about the variances, they are assumed equal, and, by the symmetry of the problem,

$$n_{T1} = n_{T2} = n_{T3}.$$

Thus, it is equivalent to minimize the sum

$$S' = 3\left(\frac{1}{n_{T1}} + \frac{1}{n_C}\right)^{1/2}$$

subject to $3n_{T1} + n_C = N$. Readily obtainable analytically using a Lagrange multiplier, the solution is given by

$$n_C = n_{T1}\sqrt{3} = n_{T2}\sqrt{3} = n_{T3}\sqrt{3}.$$

4. EXAMPLE 2

Now suppose one is designing a sample in which there are three treatment groups, one (T1) of which is matched to a comparison group (C1) and the other two (T2 and T3) of which are matched to a second comparison group (C2). Suppose the three mean differences

$$\mu_{T1} - \mu_{C1} \qquad \mu_{T2} - \mu_{C2} \qquad \mu_{T3} - \mu_{C2}$$

are to be estimated and that precision is of equal value for each.

Again assume that the differences are estimated as the differences between sample means and that the costs per unit of observation and the variances are equal among the groups. Since by the symmetry of the problem $n_{T2} = n_{T3}$, the proposed procedure then reduces to a minimization of the sum

$$S' = \left(\frac{1}{n_{T1}} + \frac{1}{n_{C1}}\right)^{1/2} + 2\left(\frac{1}{n_{T2}} + \frac{1}{n_{C2}}\right)^{1/2}$$

subject to

$$N = n_{T1} + 2n_{T2} + n_{C1} + n_{C2}.$$

Using a Lagrange multiplier, one finds the optimum allocation to be given as the solution of the following:

$$\frac{1}{2}\left(\frac{1}{n_{T1}} + \frac{1}{n_{C1}}\right)^{-1/2} n_{T1}^{-2} = \lambda$$

$$\frac{1}{2}\left(\frac{1}{n_{T1}} + \frac{1}{n_{C1}}\right)^{-1/2} n_{C1}^{-2} = \lambda$$

$$\left(\frac{1}{n_{T2}} + \frac{1}{n_{C2}}\right)^{-1/2} n_{T2}^{-2} = 2\lambda$$

$$\left(\frac{1}{n_{T2}} + \frac{1}{n_{C2}}\right)^{-1/2} n_{C2}^{-2} = \lambda$$

$$N = n_{T1} + 2n_{T2} + n_{C1} + n_{C2}.$$

From the first two equations, one sees immediately that $n_{T1} = n_{C1}$ and, similarly, from the second two, that $n_{T2} = \frac{1}{2} n_{C2}$. Since, as we have already noted, $n_{T2} = n_{T3}$, one finds the optimal allocation to be as follows:

$$n_{T1} = n_{C1} = .179\ N$$

$$n_{T2} = n_{T3} = .188\ N$$

$$n_{C2} = .266\ N$$

5. EXAMPLE 3

In a study of graduate medical education, a sample of 50 hospitals was to be selected from which data would be collected to provide estimates of means for several variables for teaching hospitals and estimates of differences in means for those same variables between teaching and nonteaching hospitals. Letting n_T and n_N denote the sample sizes for teaching and nonteaching hospitals, respectively, the optimization of the allocation of the sample between these was formulated as a problem of minimizing

$$S = w_1\left(\frac{1}{n_T}\right)^{1/2} + w_2\left(\frac{1}{n_T} + \frac{1}{n_N}\right)^{1/2}$$

subject to $n_T + n_N = 50$.

An analytic solution could not be readily found, but enumeration provided a ready and inexpensive solution. The resulting optima are shown here as a function of the ratio of w_1 to w_2:

w_1/w_2	n_T	n_N
1/4	27	23
1/3	28	22
1/2	29	21
1	31	19
2	34	16
3	36	14
4	37	13

It is noteworthy that the optimal allocation is relatively insensitive to the choice of weights.

6. EXAMPLE 4

Now suppose one is interested in estimating an overall mean and subpopulation means for two of three strata based on a sample of size 50 with relative importance weights as shown below:

Stratum	Relative Frequency	Mean to be Estimated	Relative Importance Weight
1	1/3	μ_1	.3
2	1/3	μ_2	.2
3	1/3	μ_3	0
Overall	-	$\bar{\mu}$	.5

Assuming the variances and unit costs are equal among the strata and ignoring finite population corrections, one forms the weighted sum of the standard deviations of the estimators of the means

$$S = .5\left(\frac{1}{9}\left(\frac{1}{n_1} + \frac{1}{n_2} + \frac{1}{n_3}\right)\right)^{1/2} + .3\left(\frac{1}{n_1}\right)^{1/2} + .2\left(\frac{1}{n_2}\right)^{1/2}$$

to be minimized subject to the sample size constraint $n_1 + n_2 + n_3 = 50$.

The optimal solution is readily obtainable using direct enumeration. The actual number of cases enumerated can be substantially reduqed by utilizing the following constraints:

$$n_3 \leq N/3 \quad \text{and} \quad n_3 \leq n_2 \leq N/2 .$$

The first two inequalities must hold since no estimate is needed for the third stratum mean, whereas estimates are needed for both of the other strata. Further, n_2 must be no greater than one-half N, since the relative importance for the first stratum mean exceeds that for the second stratum mean and consequently more of the sample must be allocated to the first stratum than to the second. The resulting optimal allocation is as follows:

$$n_1 = 22 \qquad n_2 = 18 \qquad n_3 = 10.$$

REFERENCES

Aoyama, H. (1963). Stratified random sampling with optimum allocation for multivariate population. Ann. Inst. Statist. Math 14, 251-258.

Bean, J.A., and Burmeister, L.F. (1978). A review of optimal sample allocation for multipurpose surveys. Biom. J. 20, 3-14.

Chakravarti, I.M. (1955). On the problem of planning a multistage survey for multiple correlated characters. Sankhya 14, 211-216.

Chatterjee, S. (1967). A note on optimum stratification. Skand. Akt., 50, 40-44.

Chatterjee, S. (1968). Multivariate stratified surveys. J. Amer. Statist. Assoc. 63, 530-534.

Cochran, W.G. (1963). Sampling Techniques. John Wiley & Sons, New York.

Dalenius, T. (1953). The multivariate sampling problem. Skand. Aktuartidskr. 36, 93-102.

Dalenius, T. (1957). Sampling in Sweden. Almquist & Wiksell, Stockholm.

Ghosh, S.P. (1958). A note on stratified random sampling with multiple characters. Calcutta Statist. Assoc. Bull. 8, 81-89.

Hartley, H.O. (1965). Multiple purpose optimum allocation in stratified sampling. Proc. Social Statist. Sect. Amer. Statist. Assoc., 258-261.

Kish, L. (1976). Optima and proxima in linear sample designs. J. Roy. Statist. Soc., Ser. A 139, 80-95.

Peters, J.H.. and Bucher, M.L. (undated). The 1940 section sample surveys of crop acreages in Indiana and Iowa. (Bureau of Agriculture Economics, U.S. Department of Agriculture.)

Yates, R. (1953, rev. 1960). Sampling Methods for Censuses and Surveys. Hafner Publishing Co., New York.