SAMPLE ALLOCATION IN MULTIVARIATE STRATIFIED DESIGN : AN ALTERNATIVE TO CONVEX PROGRAMMING

M. A Rahim, Wisner Jocelyn, Statistics Canada
M.A Rahim, 44 Birchview Road, Ottawa, K2G-3G6

ABSTRACT

In complex surveys using stratified design involving large number of variables, a common problem is how to allocate samples in different strata such that the sampling errors of estimates of population totals or averages do not exceed certain preassigned upper bounds. This problem has been addressed by recent advances in computer algorithms based on iterative procedure known as convex programming. However, Convex programming is suitable for moderately sized problems. When number of variables and strata are large, it has certain disadvantages. An alternative approach is to define a distance function of the sampling errors of all the estimates. In this paper, the performance of such a distance function is investigated using census data on clothing industry collected by Statistics Canada. The results are presented and compared with the allocation obtained under convex programming.

KEY WORDS : Complex Survey, Iterative Procedure, Distance function.

1. INTRODUCTION

In complex surveys involving large number of response variables, sometimes a stratified random sampling design is preferred. In the univariate case, sample allocation is termed as optimum in two senses. First : if the cost of the survey is preassigned, sample allocation is optimum if it minimizes the sampling error of the estimate of the population total on average. Second : if upper bound to the sampling error is preassigned, the allocation is optimum if it minimizes the survey cost. In the case of multiple response variables, following this analogy, we assign upper bounds to each of the sampling errors of estimates and then term the allocation as optimum if it minimizes survey cost. If however, the survey cost is preassigned, which is usually the case in most surveys involving multiple variables, there is no easy way to define which allocation is optimum. because there cannot be a unique set of minimum variances of the multiple estimates that we may wish to attain. To overcome this difficulty an aggregate measure of variabilities of all the estimates, in terms of a distance function of the coefficients of variation, has been proposed by Rahim and Currie (1993). This approach has been advocated from three standpoints. First : when survey cost is preassigned it is easy to define the allocation as optimum if this aggregate measure of variability is minimum. Second : if the survey involves a large number of response variables, there is not much point in being overconcerned that each of the individual variance constraints should be satisfied. For the sake of simplicity we might agree that it is good enough for all practical purposes as long as the aggregate measure of variabilities does not exceed its preassigned upper bound. Third : recently, Bethel (1989) has provided an improved algorithm for convex programming where he comments " The convex programming approach gives the optimum allocation to the defined problem but the resulting cost may not be acceptable so a further search is usually required for an optimal solution....". This in turn requires increasing the upper bounds to the variance constraints which may not be acceptable either. Sample allocation by minimizing the distance function will not lead to such an anomalous situation because the resulting survey cost is expected to be substantially lower and therefore acceptable. This third stand point, in particular, has not yet ben verified. In this paper we investigate the performance of the aggregate measure and compare with that of convex programming using census data on clothing industry collected by Statistics Canada.

2. MATHEMATICAL PRELIMINARIES

In a stratified random sampling design, let $\bar{y}_j$ denote the estimate of the population average $\bar{Y}_j$ of a variable Y_j; $j = 1, 2, \ldots, J$ and S_{ij}^2 the variance of Y_j in stratum i; $i = 1, 2, \ldots, I$. Sample and population sizes are $n = \sum_i n_i$ and $N = \sum_i N_i$. It is known that the variance of $\bar{y}_j$ is given by

$$(1) \qquad V(\bar{y}_j) = \sum_i \frac{N_i^2 S_{ij}^2}{N^2 n_i} - \sum_i \frac{N_i^2 S_{ij}^2}{N^2}$$

Bethel (1989) neglected the second term in (1) and wrote the approximate expression for $V(\bar{y})$ as

$$(2) \qquad V(\bar{y}_j) = \sum_i \frac{N_i^2 S_{ij}^2}{N^2 n_i}$$

In order to make our results comparable to his, we do the same. In that case we can write the coefficient of variation $CV(\bar{y})$ as

$$(3) \qquad CV(\bar{y}_j) = \frac{\sqrt{\sum_i \frac{N_i^2 S_{ij}^2}{N^2 n_i}}}{\bar{\bar{Y}}_j}$$

The weighted distance function of the sampling errors of $\bar{y}_j$; is defined as

$$D^2 = \sum_j W_j CV^2(\bar{y}_j) = \sum_j W_j \sum_i \frac{N_i^2 S_{ij}^2}{N^2 n_i \bar{Y}_j^2}$$

Where W_j denotes weight representing importance of the variable y_j. The individual sampling error constraints $CV(\bar{y}_j) \le \mu_j$; would imply a constraint on D^2 as

$$(4) \qquad \sum_j W_j \sum_i \frac{N_i^2 S_{ij}^2}{N^2 n_i \bar{Y}_j^2} \le D_o^2$$

Where $\qquad D_o^2 = \sum_j W_j \mu_j^2$

Writing

$$A_{ij} = \frac{N_i^2 S_{ij}^2}{N^2 n_i \bar{Y}_j^2 D_o^2}$$

and

$$x_i = \frac{1}{n_i}$$

the constraint at (4) can be written as

$$(5) \qquad \sum_j \sum_i W_j A_{ij} x_i \le 1$$

The usual cost function is written as

$$(6) \qquad g(x) = \sum_i \left(\frac{c_i}{n_i} \right) = \sum_i c_i x_i$$

where c_i is the cost of enumeration per unit in the i-th stratum. Our problem now reduces to minimizing $g(x)$ with respect to x_i subject to constraint (5). This is equivalent to minimizing a function F where

$$F = \sum_i c_i x_i + \lambda \left\{ \sum_j \sum_i W_j A_{ij} x_i - 1 \right\}$$

with respect to x_i and λ, the λ being the Lagrange multiplier. We therefore solve the equations

$$(7) \qquad \frac{\delta F}{\delta x_i} = -\frac{c_i}{x_i^2} + \lambda \sum_j W_j A_{ij} = 0$$

$$(8) \qquad \frac{\delta F}{\delta \lambda} = \sum_j \sum_i W_j A_{ij} x_i - 1 = 0$$

and obtain

$$\lambda = \left\{ \sum_i \sqrt{c_i \sum_j W_j A_{ij}} \right\}^2$$

Substituting this value of λ in (7) we obtain the formula for sample allocation as

$$(9) \quad x_i = \frac{\sqrt{c_i}}{\sqrt{\sum_j W_j A_{ij}} \sum_i \sqrt{c_i \sum_j W_j A_{ij}}}$$

For the sake of brevity, in what follows we will refer to allocations by the formula (9) as simply "D^2-Allocation" and the allocations by convex programming simply as "Convex-Allocation".

3. EMPIRICAL STUDY OF THE CONVEX-ALLOCATION AND D^2-ALLOCATION

We used Statistics Canada's census data on clothing industry. Fourteen variables Y_j; $j = 1, 2, \ldots, 14$ represent values of shipments of certain categories of clothing. These values were recorded for manufacturing industrial units classified into 9 strata by a combination of provinces of Quebec, Ontario, and certain revenue classes. We used Bethel's (1989) algorithm for convex programming setting upper bound to the coefficient of variation of the estimates $\bar{y}_j$ at 0.051. For the same data set and with the same upper bound at 0.051 we obtained allocation using the formula at (9). We assigned equal weights to all the variables. The results are shown in Table 1.

TABLE 1

Strata	Population size	Convex-Allocation	D^2-Allocation
1	216	76	51
2	342	141	80
3	512	175	122
4	467	210	175
5	482	210	114
6	100	33	30
7	107	58	35
8	369	180	153
9	343	166	126
Total	3338	1249	888

Based on the allocations shown in table 1 the coefficients of variation of the estimates $\bar{y}_j$ were computed. These are shown in table 2.

TABLE 2

Variables	Coefficients of variation of the estimates $\bar{y}_j$	
	Under Convex-Allocation	Under D^2-Allocation
1	0.044	0.052*
2	0.032	0.037
3	0.043	0.050
4	0.032	0.038
5	0.026	0.031
6	0.030	0.034
7	0.043	0.050
8	0.043	0.055*
9	0.042	0.047
10	0.043	0.049
11	0.035	0.043
12	0.029	0.033
13	0.043	0.047
14	0.025	0.030

* Cases when Cv's exceeded the upper bound

It can be seen from these tables that the total sample size required under the convex allocation is 1248 which is 37% of the population size. This shows that the cost of survey would be rather high and may not be acceptable. On the other hand, under the D^2-allocation the total sample size required is 888 which is 27% of the population size. Thus the cost of survey would be less and therefore acceptable. However, under the latter procedure coefficients of variation of 2 estimates exceed the upper bound of 0.051. In view of the simplicity of the procedure and lower cost of survey this can be regarded as an acceptable trade-off. Also it has to be stressed that the D^2-allocation procedure is intended when we have a very large number of variables and we are not that concerned about individual sampling error constraint as long as the average of the CV's of all the estimates does not exceed its preassigned upper bound. In this case the average comes out to be 0.043 which is less than the assigned upper bound 0.051. Thus our primary objective is satisfied. As an additionnal benefit, it is found that as many as 12 out of 14 individual constraints are also satisfied hence, the D^2-allocation seems to be quite a good alternative to the convex programming.

4. DISCUSSION AND SUMMARY

In stratified random sampling design involving very large number of response variables, convex programming for sample allocation to different strata has certain disadvantages. Because the procedure is quite complicated and to ensure that all the individual variance constraints are satisfied the total sample size required, and therefore the survey cost, often becomes too high and unacecceptable. Therefore, in a recent paper Rahim and Currie (1993) proposed an alternative approach based on minimizing a function D^2 which is an aggregate measure of variabilities of all the estimates of population means or totals. In this paper, we have investigated and compared the convex-allocation and the D^2-allocation using clothing industry census data of Statistics Canada. We find the performance of the D^2-allocation to be quite promising particulary from the points of view of its simplicity, lower cost of survey, relatively few violations of individual sampling error constraints, and above all it ensures the aggregate measure of sampling errors below the assigned upper bound. This last feature is important because with large number of response variables we are interested only to control the sampling errors of estimates in an aggregate sense.

REFERENCES

BETHEL, J.W. (1989). Sample Allocation in Multivariate Surveys. Survey Methodology vol 15, No 1, 47-57

CHATTERJEE, S. (1968). Multivariate Stratified Surveys. Journal of the American Statistical Association, 63, 530-534

KOKAN, A.R. and KHAN, S. (1967). Optimum Allocation in Multivariate Surveys: An Analytical Solution. Journal of the Royal Statistical Society B., 29, 115-125

RAHIM, M.A. and CURRIE, S. (1993). Optimizing Sample Allocation with multiple Response Variables. Proceedings of the American Statistical Association: Survey Research Methods Section.