OPTIMIZING SAMPLE ALLOCATION FOR MULTIPLE RESPONSE VARIABLES

M.A. Rahim, S. Currie, Statistics Canada
M.A. Rahim, 44 Birchview Road, Ottawa, Ontario

KEY WORDS: *Convex programming, Stratified random sampling, Neyman allocation.*

ABSTRACT

In a stratified random sampling design involving multiple response variables, convex programming optimizes sample allocation in the sense that the preassigned upper bounds to the sampling errors of the estimates are satisfied while minimizing the cost. Its use, however, is limited to moderately sized problems only. The other optimality criterion is not addressed either, namely, if the cost is preassigned, how to determine the allocation that will minimize, in some sense, the sampling errors of all the estimates. Based on a distance function, measuring aggregate variabilities of the estimates, we provide a solution either when the cost is preassigned or an upper bound to the aggregate measure is preassigned. It is shown that the Neyman Allocation is a special case to our extended formula for multiple variables. We also investigate a simple alternative to convex programming under which, when large number of response variables are involved, not all, but a good majority of the sampling error constraints are likely to be satisfied. Results of the investigation, using actual Statistics Canada data are presented.

1. INTRODUCTION

For the purpose of estimating averages or totals of the values of multiple response variables, sometimes a simple stratified random sampling design is preferred. In that case, following univariate analogy, sample allocation to different strata can be qualified as optimum in two senses. First: If the cost of the survey is preassigned, the allocation should minimize, in some sense, the sampling errors of all the estimates. Second: If upper bounds to the sampling errors of the estimates are preassigned, the allocation should minimize the cost. In fact, there is no exact solution to this problem, although in the latter case optimization is possible by an iterative procedure, namely, convex programming. Recently, Bethel (1989) has provided an improved algorithm for the purpose but points out a number of practical difficulties (p. 47, 53). First: "The convex programming approach gives the optimal solution to the defined problem but the resulting cost may not be acceptable so a further search is usually required for an optimal solution . . .". And this has to be done obviously - by increasing the upper bounds to the variance constraints that we had originally set. Second: "The algorithm converges quickly for most moderately sized problems" and the "run times vary considerably depending on the magnitude of the problem . . .". Third: " . . . labour involved in creating files and other preparatory tasks "is of much greater concern than merely the run times". A further point to be noted is the fact that in practice, usually the cost of the survey is preassigned. And in that case the optimality criterion has not been addressed, namely, how to determine the allocation that will minimize, in some sense, the sampling errors of all the estimates.

The problem is further aggravated by the fact that in actual surveys quite a large number of response variables are involved. A case, for example, is the annual sample survey of manufacturing establishments by Statistics Canada, for the purpose of estimating thousands of commodity outputs. In such cases, convex programming is impractical. One option to resolve this problem is to define optimum allocation in terms of some aggregate measure of variabilities of all the estimators.

In this paper we propose a distance function, measuring aggregate variabilities of the estimates and provide solutions, either in the case when cost is preassigned or in the case when an upper bound to the aggregate measure is preassigned. Under this formulation, the Neyman allocation is shown to be a special case to our extended formula for multiple response variables.

We also investigate another possibility, namely, with very large number of response variables we may not be overconcerned to ensure that every

individual variance constraint is satisfied. For the sake of simplicity we may rather accept a sample allocation procedure under which, if not all, a good majority of the variance constraints are satisfied. Such a procedure is illustrated using actual census data on clothing industry from Statistics Canada. The result appears to be satisfactory for all practical purposes.

2. THE PROBLEM

In a stratified random sampling design, let $\hat{X}_j$ denote the estimate of the total value X_j of a variable x_j; $j=1,2,\ldots,p$; and S_{jh}^2 the variance of x_j; in stratum h; $h=1,2,\ldots,L$. Sample and population sizes are $n = \sum_h n_h$ and $N = \sum_h N_h$ respectively. Writing $a_{jh}^2 = N_h^2 S_{jh}^2 / N^2$ as the weighted stratum variance, we can write the variances of the estimates and the cost function as

$$V(\hat{X}_j) = \sum_h (N a_{jh})^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right); \quad j = 1,2,\ldots,p \tag{2.1}$$

$$c = \sum_h n_h c_h \tag{2.2}$$

where c_h is the cost of enumeration per unit in stratum h. For any single variable x_r the optimum allocation (Neyman 1934, Tschuprow 1923) can be written in equivalent form as

$$n_h = \frac{(c_0 a_{rh}) / \sqrt{c_h}}{\sum_h a_{rh}\sqrt{c_h}}; \quad h = 1,2,\ldots,L \tag{2.3}$$

for a preassigned cost $c = c_0$, and

$$n_h = \frac{a_{rh}/\sqrt{c_h} \sum_h a_{rh}\sqrt{c_h}}{(v_r/N^2) + \sum_h (a_{rh}^2/N_h)}; \quad h = 1,2,\ldots,L \tag{2.4}$$

for a preassigned upper bound to the sampling error of the estimate $V(\hat{X}_r) = v_r$. It is obvious from (2.3) and (2.4) that in either case what is optimum allocation for x_r, will not be optimum for x_k simply because $a_{rh} \neq a_{kh}$; $k \neq r = 1,2,\ldots,p$.

One option to resolve this problem is to define a distance function of the sampling errors of the estimates. In actual surveys the sampling error is usually expressed in terms of the coefficient of variation cv. We therefore choose a distance function D of the form

$$D = W_1 cv^2(\hat{X}_1) + W_2 cv^2(\hat{X}_2) + \ldots + cv^2(\hat{X}_p) \tag{2.5}$$

where W_j is a weight reflecting the importance of the variable x_j. Note that it makes sense to use D as an aggregate measure of the variabilities of all the estimators $\hat{X}_j$; $j = 1,2,\ldots,p$. We thus consider optimizing the sample allocation in terms of this aggregate measure D.

3. SOLUTION

3.1 WHEN COST IS PREASSIGNED

From (2.1) we can write $cv^2(\hat{X}_j)$ as

$$cv^2(\hat{X}_j) = \sum_h (N a_{jh}/X_j)^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right); \quad j = 1,2,\ldots,p \tag{3.1}$$

Substituting these in (2.5) we get

$$D = \sum_j \sum_h \frac{W_j A_{jh}^2}{n_h} - \sum_j \sum_h \frac{W_j A_{jh}^2}{N_h} \tag{3.2}$$

where we write $A_{jh} = (N a_{jh}) / X_j$. Our objective now is to minimize D subject to the condition $c = \sum_h n_h c_h = c_0$. This is equivalent to minimizing the function F where

$$F = \sum_j \sum_h \frac{W_j A_{jh}^2}{n_h} - \sum_j \sum_h \frac{W_j A_{jh}^2}{N_h} + \lambda \left(\sum_h n_h c_h - c_0\right) \tag{3.3}$$

and λ is the Lagrange multiplier. Differentiating F with respect to n_h; $h = 1,2,\ldots,L$; and λ and equating to zero we have

$$\frac{\delta F}{\delta n_h} = -\sum_h \frac{W_j A_{jh}^2}{n_h^2} + \lambda c_h = 0; \quad h = 1,2,\ldots,L \tag{3.4}$$

$$\frac{\delta F}{\delta \lambda} = \sum_h n_h c_h - c_0 = 0 \qquad (3.5)$$

Solving these equations we get $\lambda = \left[\left(\sum_h \sqrt{\sum_j c_h W_j A_{jh}^2}\right)/c_0\right]^2$ and substituting this value of λ in (3.4) the desired allocation for multiple variables is obtained as

$$n_h = \frac{c_0 \sqrt{\left(\sum_j W_j A_{jh}^2\right)/c_h}}{\sum_h \sqrt{\left(\sum_j W_j A_{jh}^2\right) c_h}} \; ; \qquad (3.6)$$
$$j = 1, 2, \ldots, p$$
$$h = 1, 2, \ldots, L$$

In the case of any single variable x_r, since $A_{rh} = Na_{rh}/X_r$, the above reduces to

$$n_h = \frac{(c_0 a_{rh})/\sqrt{c_h}}{\sum_h a_{rh}\sqrt{c_h}} \; ; \qquad (3.7)$$
$$h = 1, 2, \ldots, L$$

which is same as the Neyman Allocation in (2.3).

If $c_h = c$ i.e., per unit cost is same in all strata then we have $a_{rh} = (N_h S_{rh}) / N$. Also we have $c_0 = \sum_h n_h c_h = nc$. Hence, (3.7) can also be written as

$$n_h = \frac{N_h S_{rh}}{\sum_h N_h S_{rh}} * n; \qquad (3.8)$$
$$h = 1, 2, \ldots, L$$

which is the more familiar form of Neyman Allocation when n is known. Thus Neyman Allocation is a special case to our extended formula (3.6) for multiple variables.

3.2 WHEN AN UPPER BOUND TO THE AGGREGATE MEASURE OF THE SAMPLING ERRORS IS PREASSIGNED

Let the individual sampling error constraints be written as $CV(\hat{X}_j) \le \mu_j ; j = 1, 2, \ldots, p$. If large numbers of variables are involved, it is assumed that we are concerned only with an upper bound to the aggregate measure of the sampling errors which, in this case, is $D_0 = \sum_j W_j \mu_j^2$. Hence, our problem is to minimize the cost $C = \sum_h n_h c_h$ subject to the condition $D = \sum_j W_j CV^2(\hat{X}_j) \le D_0$.

This is equivalent to minimizing the function F where

$$F = \sum_h n_h c_h + \lambda \left[\sum_j \sum_h \frac{W_j A_{jh}^2}{n_h} - \sum_j \sum_h \frac{W_j A_{jh}^2}{N_h} - D_0 \right] \qquad (3.9)$$

and λ is the Lagrange multiplier. Differentiating F with respect to n_h ; $h = 1, 2, \ldots, L$; and λ and equating to zero we have

$$\frac{\delta F}{\delta n_h} = c_h - \lambda \sum_j \frac{W_j A_{jh}^2}{n_h^2} = 0 ; \qquad (3.10)$$
$$h = 1, 2 \ldots, L$$

$$\frac{\delta F}{\delta \lambda} = \sum_j \sum_h \frac{W_j A_{jh}^2}{n_h} - \sum_j \sum_h \frac{W_j A_{jh}^2}{N_h} - D_0 = 0 \qquad (3.11)$$

Solving these equations we get

$$\lambda = \frac{\left[\sum_h \sqrt{c_h}\left(\sqrt{\sum_j W_j A_{jh}^2}\right)\right]^2}{\left[D_0 + \sum_j \sum_h (W_j A_{jh}^2)/N_h\right]^2}$$

and substituting this value of λ in (3.10) the desired allocation for multiple variables is obtained as

$$n_h = \frac{\left(\sqrt{\sum_j W_j A_{jh}^2}/\sqrt{c_h}\right)\left(\sum_h \sqrt{c_h}\sqrt{\sum_j W_j A_{jh}^2}\right)}{D_0 + \sum_j \sum_h (W_j A_{jh}^2)/N_h} ; \qquad (3.12)$$
$$h = 1, 2, \ldots, L$$

In the case of a single variable x_r - noting that $A_{rh} = Na_{rh}/X_r$ and D_0 reduces to $W_r \mu_r^2 = W_r, v_r / X_r^2$ where $V(\hat{X}_r) = v_r$ is the preassigned upper bound in terms of the variance - the above

reduces to

$$n_h = \frac{(a_{rh}/\sqrt{c_h})\sum_h a_{rh}\sqrt{c_h}}{(v_r/N^2) + \sum_h (a_{rh}^2/N_h)}; \quad (3.13)$$

$$h = 1, 2, \ldots, L$$

which is same as the Neyman Allocation in (2.4). Thus, the Neyman Allocation in the case of a preassigned upper bound to the sampling error of the estimate, is also a special case to our extended formula (3.12) for multiple variables.

4. EMPIRICAL STUDY

In this section we present the results of a very simple allocation procedure which may be useful when number of variables are large, no distinction is made about their relative importance, and we are not overconcerned that each individual sampling error constraint must be satisfied.

Using Neyman Allocation for single variables we can get the allocation for each of the p variables in stratum h as $n_{1h}, n_{2h}, \ldots, n_{ph}$. Let us order them and denote the maximum one i.e., the 100-th percentile value as $n_{h(100)}$. We may take $n_{h(100)}; h = 1, 2, \ldots, L$ as the required allocation for which, obviously, all the p constraints $cv(\hat{X}_j) \leq \mu_j; j = 1, 2, \ldots, p$ will be satisfied. But the total sample size $n = \sum_h n_{h(100)}$ would be very large and therefore the cost would be prohibitive.

We can reduce the cost i.e., total sample size n by selecting another allocation $n_{h(p)}$; $h = 1, 2, \ldots, L$ with some $p < 100$. But then all of the constraints $cv(\hat{X}_j) \leq \mu_j; j = 1, 2, \ldots, p$ would not be satisfied. Hence, the crucial question is, can we choose a suitable p such that only a small number of constraints are not satisfied and at the same time the cost is acceptable? If so, then it makes sense - just for the sake of simplicity - to conduct the survey based on the allocation $n_{h(p)}$; $h = 1, 2, \ldots, L$.

It is this question we investigated relating to 14 commodity outputs of the clothing industry during the year 1986 for which complete census data were available from Statistics Canada. The population consisted of N = 2256 manufacturing establishments. There were 39 strata based on combination of Province and revenue class. We set the upper bounds to the sampling error of estimates of each of the 14 commodities at 10%, i.e., we targeted $cv(\hat{X}_j) \leq 10\%$ where $\hat{X}_j$ represents the estimate of the j-th commodity output; $j = 1, 2, \ldots, 14$. Using three different allocations for P = 100, 75, 50 (i.e., maximum , 75th percentile, and median) the sampling errors of estimates and also the total sample size n, required under each allocation were found to be as shown in Table 1.

TABLE 1: Values of the coefficient of variation of each of the 14 commodity output estimates along with the total sample size n under three different allocations.

	p=100	p=75	p=50
1	0.038	0.103*	0.193*
2	0.009	0.030	0.047
3	0.020	0.060	0.085
4	0.035	0.093	0.131*
5	0.025	0.068	0.150*
6	0.015	0.046	0.065
7	0.020	0.051	0.069
8	0.053	0.239*	0.338*
9	0.030	0.073	0.101*
10	0.026	0.079	0.125*
11	0.027	0.062	0.086
12	0.061	0.151*	0.257*
13	0.029	0.089	0.124*
14	0.031	0.091	0.145*
	n=1139	n=680	n=570

* Indicates cv higher than the preassigned value 10%

It is found - as it should - that for $p = 100$ all the cvs are less than the preassigned upper bound of 10% but the total sample size $n = 1139$ is large for a population of $N = 2256$ units. For $p = 50$, sample size is small but as many as 9 cvs exceed 10%. For $p = 75$, however, the sample size $n = 680$ is reasonable and only 2 cvs significantly exceed 10%.

We then investigated that for this $n=680$ (i.e., fixed cost) how the minimum average cv attainable by using the extended allocation formula (3.6) compares with what we obtained based on $p=75$. Table (4.2) below shows the comparative results.

Table 4.2 Values of the coefficient of variation of the estimates under extended Neyman allocation and the allocation based on $p=75$.

	p=75	under extended Neyman Allocation
1	0.103	0.103
2	0.030	0.029
3	0.060	0.058
4	0.093	0.093
5	0.068	0.079
6	0.046	0.044
7	0.051	0.048
8	0.239	0.201
9	0.073	0.072
10	0.079	0.075
11	0.062	0.060
12	0.151	0.152
13	0.089	0.084
14	0.091	0.089
Average cv	0.008	0.085

It is found that the average cv 0.088 under the allocation based on $p=75$ came out to be almost same as the minimum average cv 0.085 attainable for a fixed $n=680$. In other words, given the cost required for a sample size of 680 we would not get any better allocation other than what we got for $p=75$ from the point of view of minimizing the average sampling error.

In view of the above findings it seems that when we are confronted with the task of a large number of estimations, and convex programming becomes virtually impossible, such a simple allocation procedure based on a suitable choice of percentile (p) value can be used as a compromise solution.

5. DISCUSSION AND SUMMARY

This study addresses the problem of optimum allocation of sample sizes to different strata in a stratified random sampling design with multiple response variables - particularly when the number of variables are large. We propose an optimality criterion based on a weighted distance function D of the sampling errors of estimates of the totals of all the response variables under study. Based on this distance function, which is in fact an aggregate measure of variability of all the estimates, we provide a solution either in the case when cost of the survey is preassigned or in the case when an upper bound to the aggregate measure is preassigned. It is shown that the well known Neyman allocation in either situation, is a special case. It should be noted, however, that the allocations - as can be seen from (3.6) and (3.12) - depends on a set of weights $W_j ; j = 1, 2, \ldots, p$ assigned arbitrarily to reflect the importance of the variables $x_j ; j = 1, 2, \ldots, p$. Usually a rare or unimportant variable will have larger variability in the population. If we are willing to assume that the more the variability of a variable x_j; in the population, the less is its importance, then it makes sense to choose the weights as $W_j = 1 / \hat{V}(\hat{X}_j)$ where $\hat{V}(\hat{X}_j)$ is the sample estimate of $V(\hat{X}_j)$. It has also been pointed out that convex programming - although mathematically attractive - becomes an impractical proposition in actual surveys when we have to estimate a large number of variable values. This justifies a search for some alternative procedure whose simplicity can be traded off with the limitation that a few of the sampling error constraints need to be violated. Such a procedure has been illustrated using census data on the clothing industry from Statistics Canada. The result, admittedly a compromise solution, appears to be satisfactory for all practical purposes.

ACKNOWLEDGEMENT

The authors wish to acknowledge receiving considerable help in programming and computation from Danielle Lalande, Wisner Jocelyn, and Douglas Yeo of the Business Survey Methods Division of Statistics Canada. Thanks are also due

to Linda Lafontaine and Carole Jean-Marie who meticulously typed this paper.

REFERENCES

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

Bethel, J.W. (1985), "An Optimum Allocation Algorithm for Multivariate Surveys," Proceedings of the Survey research section, American Statistical Association, 209-212.

Chatterjee, S. (1972), "A Study of Optimum Allocation in Multivariate Stratified Surveys," *Skandinavisk Actuarietidskrift*, 55, 73-80.

Cochran, W.G. (1953), *Sampling Techniques*, New York: John Wiley.

Dalenius, T. (1953), "The Multivariate Sampling Problem," *Skandinavisk Actuarietidskrift*, 36, 92-102.

Folks, J.L. and Antle, C.E. (1965), "Optimum Allocation of Sampling Units When There are R Responses of Interest," *Journal of the American Statistical Association*, 60, 225-233.

Hartley, H.O. (1965). "Multiple purpose optimum allocation in stratified sampling," Proceedings of the Social Statistics Section, American Statistical Association, 258-261.

Kish, L. (1976), "Optima and proxima in linear sample designs," *Journal of the Royal Statistical Society A*, 139, 80-95.

Kokan, A.R. (1963), "Optimum allocation in multivariate surveys," *Journal of the Royal Statistical Society A*, 126, 557-565.

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.

Kuhn, H.W. and Tucker, A. W. (1951), "Nonlinear Programming," Proceedings of the 2nd Berkeley Symposium Mathematical Statistics and Probability.

Tschuprow, A.A. (1923), "On the Mathematical Expectation of the Moments of Frequency Distributions in the Case of Correlated Observations," *Metron 2*, 461-493, 646-683.