CALIBRATION ALLOCATION OF SAMPLE FOR MULTIPLE CHARACTERISTIC SURVEYS UNDER STRATIFIED RANDOM SAMPLING

Avinash C. Singh and Owen Phillips[1]

ABSTRACT

This paper examines the problem of sample allocation for a stratified random design given multiple characteristics of interest. Existing solutions use non-linear programming to obtain a minimum cost and optimal allocation for a given set of variance constraints. If the resulting cost is not acceptable, the existing solution relaxes all variance constraints uniformly (i.e. allows uniform tolerance) such that an acceptable cost can be met. This is not reasonable, as not all of the constraints have an impact of the same order. An alternative formulation is required where there is a direct control on cost and the objective function is defined in terms of the variance constraints such that the two formulations become equivalent. Such a formulation is proposed using the idea of penalized distance function as in the ridge-calibration of sampling weights in Rao and Singh (1997). The proposed method is termed calibration allocation and allows for suitable, non-uniform tolerances when cost is reduced.

KEY WORDS: Directional tolerance; Optimal cost; Penalized distance function; Sample allocation;

RÉSUMÉ

Cet article examine le problème de répartition d'échantillon pour un plan d'échantillonnage stratifié simple où il y a plusieurs variables d'intérêt. Les solutions actuelles utilisent la programmation non linéaire afin d'obtenir le coût minimum et la répartition optimale pour les contraintes de variance spécifiées. Si le coût réalisé n'est pas acceptable, la solution actuelle relaxe toutes contraintes de variance d'une manière uniforme (c.-à-d. permet une tolérance uniforme) de façon à ce qu'un coût acceptable soit atteint. Ceci n'est pas raisonnable car toutes les contraintes n'ont pas un impact de la même importance. Une formulation alternative est requise là où il y a un contrôle direct sur le coût et la fonction objective est définie en termes des contraintes de variance de façon à ce que les deux formulations soient équivalentes. Une telle formulation est proposée en utilisant l'idée d'une fonction de distance pénalisée comme dans le calage-réduction des poids d'échantillonnage dans Rao et Singh (1997). La méthode proposée s'appelle la répartition par calage et permet une tolérance non uniforme appropriée lorsque le coût est réduit.

MOTS CLÉS : Coût optimal; répartition d'échantillon; fonction de distance pénalisée; tolérance directionnelle

1. INTRODUCTION AND MOTIVATION

1.1 Description of the Problem

Sample size determination is one of the primary considerations when planning a sample survey. The quality of inference that can be made about the target population is linked to the number of units sampled. As a result, a number of factors need to be considered when calculating sample size, including the desired precision of the resulting estimates, the cost of sampling, and other operational constraints.

For stratified simple random sampling (STSRS), given a single characteristic of interest, we often seek the allocation that minimizes the cost C of sampling with some upper bound v on the variance, $\hat{V}(\hat{\mu})$ or $\hat{V}(\hat{Y})$, of the estimated population mean or total of a variable Y. Conversely, if the optimal cost is too high, the equivalent problem of minimizing variance for a given cost can be solved. For a single variable of interest, simple closed-form solutions exist for the optimal allocation such that the two solutions are equivalent (see Cochran 1977, pp.96-99)
While the theory for a single characteristic is well developed and simple to implement, surveys are generally multipurpose. Given multiple characteristics of interest—and hence multiple variance constraints—the optimal cost can

1 Avinash C. Singh(asingh@rti.org), Statistics Research Division, RTI International, RTP, NC, USA 27709 and Owen Phillips(owen.phillips@statcan.ca), Statistics Canada, 11th Floor R.H. Coats Bldg., Ottawa, Ontario, Canada, K1A 0T6

be obtained using non linear programming (NLP); see, for example, Bethel (1985, 1989), Chromy (1987). However, if the optimal cost is too high, how do we relax variance constraints optimally subject to reduced cost? Earlier proposals include minimizing a linear function of variances subject to reduced cost—see, for example, Rahim and Currie (1991)—but, this approach provides no guidance as to how the individual constraints are relaxed. Bethel (1989) proposed the uniform relaxation of variance constraints, however, in doing so, some constraints are relaxed unnecessarily, and restrictions on the minimum stratum sample sizes (often imposed in practice) may no longer be satisfied. Ideally, we would like to find an equivalent formulation in terms of minimizing a function of variance constraints for a given cost for the multivariate case as in the univariate case.

1.2 Motivation from ridge-calibration of sampling weights

We take our motivation from the ridge-calibration of sampling weights, as in Rao and Singh (1997). The goal there is to minimize a chi-square distance-type function with penalty term

$$\Delta\left(\underset{\sim}{w}, \underset{\sim}{w}^0 \mid \underset{\sim}{\lambda}\right) = \sum_{k=1}^{n}\left(w_k - w_k^0\right)^2 \Big/ w_k^0 + \sum_{j=1}^{p}\left(\sum_{k=1}^{n} x_{jk} w_k - \tau_j\right)^2 \Big/ \lambda_j \qquad (1)$$

defined in terms of the adjusted and unadjusted sampling weights w_k and w_k^0, subject to benchmark constraints $\sum_{k=1}^{n} x_{jk} w_k = \tau_j$, j=1,…,J, where x_{jk} represents the value of the j^{th} auxiliary variable for the k^{th} observation. For large J, it may be impossible to satisfy exactly all benchmark constraints. The inverse penalty λ_j controls, in a sense, the amount of tolerance allowed in satisfying the benchmark constraint: as $\lambda_j \to 0$, the corresponding benchmark constraint becomes binding; as $\lambda_j \to \infty$, the corresponding constraint becomes non-binding with unlimited tolerance.

Thus, thinking of the stratum allocations as the weights and the variance constraints as the benchmark constraints in the above formulation, and defining the unadjusted 'weight' in terms of some simple known allocation—say proportional or equal—one can see how this approach might easily be adapted to the allocation of sample. There is however a problem. Tolerances in the calibration of sampling weights are symmetric with respect to $\sum_{k=1}^{n} x_{jk} w_k - \tau_j = 0$. In other words, if the benchmark constraint must be relaxed, whether we overestimate or underestimate the desired total τ_j is of little consequence as long as we are close. For sample allocation problem, however, tolerances are directional with respect to the variance constraints i.e. the desired variance is an upper bound and if the allocation of sample more than satisfies this upper bound, all the better. Thus, for the problem of multiple characteristic sample allocation, an asymmetric penalty function is required. This is what the proposed method tries to accomplish.

2. CALIBRATION ALLOCATION

2.1 The proposed method

Let N be the population size measure, H be the number of strata and J>2 be the number of characteristics of interest (variance constraints) and let W_h^2 and S_{hj}^2 represent respectively the proportion of the population in stratum h and the (estimated) stratum variance of characteristic j. For the purpose of illustration, consider the variance $V(\hat{\mu}_j)$ of the estimated population mean, j=1,…,J. We would like $\hat{V}(\hat{\mu}_j) = \sum_h W_h^2 S_{hj}^2 / n_h - \sum_h W_h S_{hj}^2 / N \le v_j$ for all j.

We can define $v_j^* = v_j + \sum_h W_h S_{hj}^2 / N$ so that variance constraints are written as $\sum_h n_h^{-1} x_{hj} - 1 < 0$, where $x_{hj} = W_h^2 S_{hj}^2 / v_j^*$. We can then define a distance-type objective function in terms of the variance constraints

$$F = \ln\left(1 + \exp\left[\lambda^{-1}\left(\sum_h n_h^{-1} x_{hj} - 1\right)\right]\right) \qquad (2)$$

where λ is an inverse penalty. The goal is to minimize F, for sufficiently small λ subject to a binding constraint $\sum_h c_h n_h = \tau$ on the cost of sampling and restrictions $L_h \le n_h \le U_h$ to the stratum sample sizes, where c_h represents the per-unit cost of sampling in stratum h. We give the name *calibration allocation* to the allocation $\{n_h\}_{h=1}^{H} = \{n_h^{cal}\}_{h=1}^{H}$ that minimizes (2) for given cost and sample size restrictions. It is assumed that there is a unique minimizer.

The above formulation differs somewhat from the objective function presented in (1) in that it consists only of a penalty term and a common inverse penalty is applied to all constraints. The first term in (1) was not used because the second term dominates. Moreover, the flexibility of a constraint specific inverse penalty was deemed unnecessary because of the nature of variance constraints, i.e., tolerance is not needed on bounds.

2.2 Properties of the objective function

Consider the exponential terms in (2) as $\lambda \to 0$: for constraints that are satisfied with strict inequality—i.e. where $\sum_h n_h^{-1} x_{hj} - 1 < 0$ —the term $\exp\left[\lambda^{-1}\left(\sum_h n_h^{-1} x_{hj} - 1\right)\right] \to 0$; for constraints that are satisfied with equality—i.e. where $\sum_h n_h^{-1} x_{hj} - 1 = 0$ —the term $\exp\left[\lambda^{-1}\left(\sum_h n_h^{-1} x_{hj} - 1\right)\right] \to 1$; and, for constraints that are not satisfied—i.e. where $\sum_h n_h^{-1} x_{hj} - 1 > 0$ —the term $\exp\left[\lambda^{-1}\left(\sum_h n_h^{-1} x_{hj} - 1\right)\right] \to +\infty$. Thus, for smaller values of λ, the penalty function is better able to distinguish between allocations where variance constraints are satisfied with strict inequality and those where they are not. Hence, minimization of F for sufficiently small values λ will result in allocations that attempt to satisfy the variance constraints as best as possible. Defining the objective function as the sum of the exponential terms would be sufficient to reflect this asymmetry; however, to alleviate the potential for problems of numeric instability, the log-transformation was made to get back to the linear scale and an offset term of 1 was added in (2) to maintain the asymmetry and positivity of the terms.

The following propositions show equivalence of calibration and optimal allocations under certain conditions.

2.3 Proposition 1 - Equivalence of optimal and calibration allocations for given variance constraints

Let the *optimal allocation* be the unique solution $\{n_h\}_{h=1}^H = \{n_h^{opt}\}_{h=1}^H$ to the cost minimization problem

Minimize $\sum_{h=1}^H c_h n_h$

Subject to $V_j < v_j$, $j = 1,...,J$ and $L_h \le n_h \le U_h$, $h = 1,...,H$.

and let $\tau^{opt} = \sum_{h=1}^H c_h n_h^{opt}$ be the optimal cost.

Then, there exists λ sufficiently small such that the calibration allocation for a cost τ^{opt} and the given variance and sample size constraints is identical to the optimal allocation. A justification of this statement is provided in the appendix.

Proposition 1 seems to present a circular argument, in that it requires the optimal cost as input to calibration allocation. However, a bisection search of a suitable cost grid could be used coupled with calibration allocation to arrive at the optimal solution. That being said, use of calibration allocation to obtain the optimal allocation is secondary—our interest lies mainly in what it can do for suboptimal costs.

2.4 Proposition 2 - Equivalence of calibration and optimal allocations for a given reduced cost

Let $\tilde{v}_j$, j=1,…,J be the realized variances under calibration allocation for a given reduced cost $\tilde{\tau} < \tau^{opt}$ and λ sufficiently small. Then, for relaxed variance constraints $\hat{V}_j \le \max\left(v_j, \tilde{v}_j\right)$, the revised optimal allocation and the corresponding optimal cost are identical to the calibration allocation and the corresponding reduced cost respectively. This can be shown using the lemma that calibration allocation for a given reduced cost $\tilde{\tau} < \tau^{opt}$ and the initial variance constraints $\hat{V}_j \le v_j$ is equivalent to calibration allocation for the given reduced cost $\tilde{\tau} < \tau^{opt}$ and the relaxed variance constraints $\hat{V}_j \le \max\left(v_j, \tilde{v}_j\right)$. See the appendix for the proof.

3. EMPIRICAL RESULTS

Data used for the purpose of this example are those presented in Bethel (1985). Results from a USDA survey of farms provide the a priori information necessary to calculate the sample sizes (i.e. the number of farms to be sampled) for either a new survey or a follow-up. There are J=9 variables of interest: number of cattle (*var1*); bushels of stored corn (*var2*); bushels of stored soybeans (*var3*); number of dairy cattle (*var4*); acres of planted corn (*var5*), soybeans (*var6*), wheat (*var7*) and hay (*var8*); and, number of hogs (*var9*). The population size N=133,574 and there are H=11 strata. The cost

of sampling a single unit in each stratum is given as $c_h = 6$ for $h = 1,...,8$ and $c_h = 140$ for $h = 9,10,11$. For additional information including the stratum sizes N_h (number of farms), stratum standard variances S_{hj}^2 and estimated population totals, the reader is referred to Bethel (1985).

For the purpose of example, we assume that a maximum coefficient of variation CV_1 of 10% is desired for the estimated mean $\hat{\mu}_1$ of *var1*, and that CV_j of 8% are desired for the estimated means $\hat{\mu}_2,...,\hat{\mu}_9$ of the remaining eight variables. Thus, upper bounds on the variances are obtained as $v_j = \hat{\mu}_j{}^2 CV_j{}^2$ for all *j*. Also, we assume loose bounds L_h=1 and U_h= N_h for all *h*.

The optimal allocation is obtained using Statistics Canada's generalized system for sampling, known as GSAM. This system uses an extension of the algorithm presented in Bethel (1989); for details, see Estevao (1998). Calibration allocations are obtained by iterative use of the SAS non-linear programming procedure PROC NLP. Decreasing values of $\lambda = 2^{1-\nu}$ are used until such a time as $\max_h \left[abs\left(n_h^{(\nu)} - n_h^{(\nu-1)}\right)\right] < \varepsilon$, for some prescribed value $\varepsilon > 0$ or numerical overflow occurs, where ν indexes the iteration. For suboptimal costs and small λ, overflow is anticipated, however it is hoped that values of the allocations between successive iterations will have stabilized significantly. A value of $\varepsilon = 10^{-4}$ was used for this example. The initial allocation is defined as $n_h^{(0)} = L_h + \left(\tau - \sum_{i=1}^{H} c_i L_i\right)\left(\sum_{i=1}^{H} c_i \left(U_i - L_i\right)\right)^{-1}\left(U_h - L_h\right)$ so that the cost constraint and bounds on sample size are satisfied.

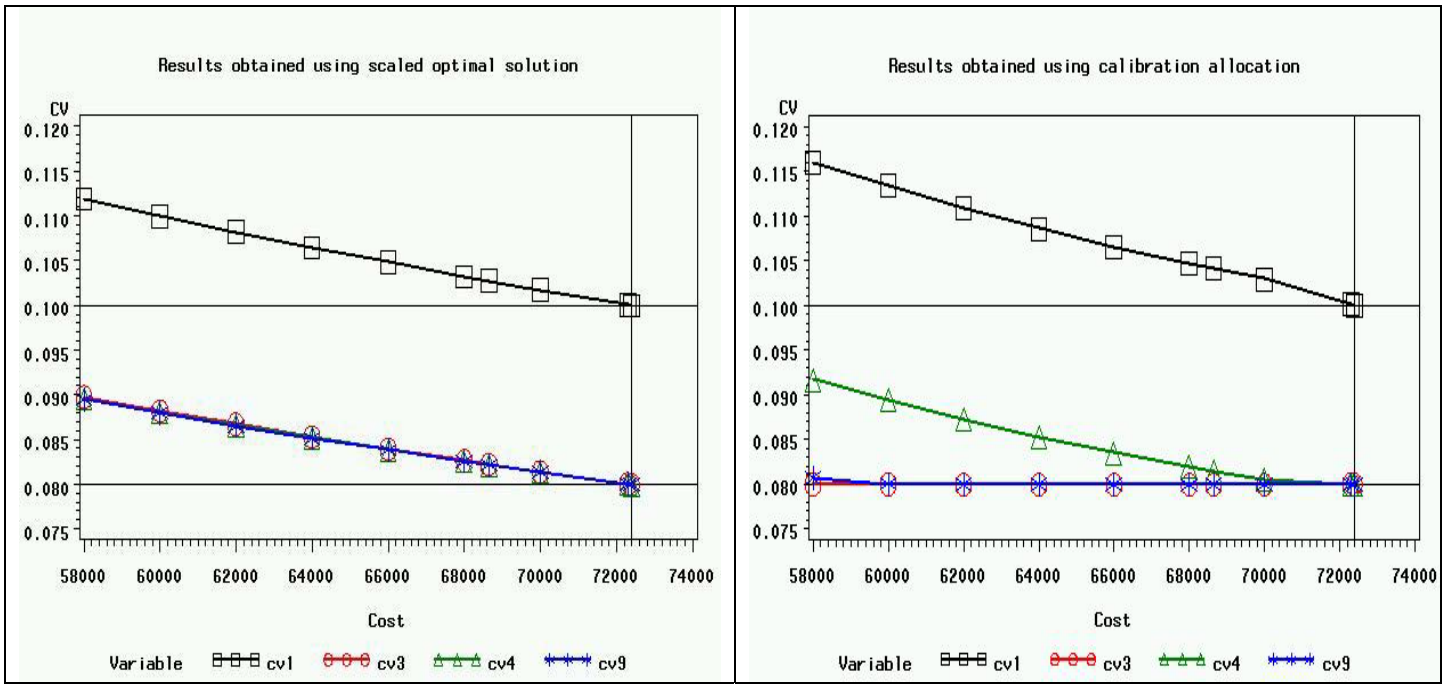
Using GSAM, the optimal sample sizes were found to be 2005.18, 78.96, 7.04, 280.64, 926.25, 168.44, 95.51, 9.68, 303.87, 26.70 and 33.47 in the eleven strata respectively, for a total cost of approximately 72,396. The expected CV's under this allocation are approximately 10.00% (*var1*), 4.54% (*var2*), 8.00% (*var3*), 8.00% (*var4*), 2.20% (*var5*), 2.40% (*var6*), 4.66% (*var7*), 4.70% (*var8*), 8.00 (*var9*) for the nine variables. Thus, it would seem that the solution is dominated by the constraints placed on the variances of the estimated means of *var1*, *var3*, *var4* and *var9*. Thus, if the optimal cost of 72,396 is too high, at least one of these constraints must be violated in order to satisfy budgetary constraints.

Using calibration allocation for a cost $\tau = 72{,}396$, nearly identical allocations of 2005.17, 78.96, 7.04, 280.64, 926.27, 168.44, 95.51, 9.68, 303.87, 26.70 and 33.47 are obtained. There are negligible differences between the allocations provided by the two approaches as expected under Proposition 1. It should be noted that, in practice, the stratum sample sizes would be rounded up to the nearest integer. However, given the binding cost constraint of calibration allocation, for the purpose of coherence decimal values are presented here.

Now consider a suboptimal cost, say $\tilde{\tau} = 58{,}000$. For the reduced cost, calibration allocation results in stratum sample sizes of 1310.11, 119.89, 6.84, 218.49, 1027.22, 212.77, 74.06, 10.00, 224.60, 36.20 and 25.80. The expected CV's under this allocation are approximately 11.60% (*var1*), 4.58% (*var2*), 8.00% (*var3*), 9.17% (*var4*), 2.33% (*var5*), 2.48% (*var6*), 4.81% (*var7*), 5.13% (*var8*), 8.07 (*var9*) for the nine variables. Using the expected CV's to revise the variance constraints, the optimal allocation produces (with negligible difference) the same allocation, providing support for Proposition 2. Note that the variance constraints on *var1*, *var3* and *var9* are violated, however all other constraints—including that on *var4*—are satisfied.

Bethel's (1989) suggestion of scaling the optimal allocation by a factor $m = \tilde{\tau}/\tau^{opt}$ results in a proportional increase to all variances. Figure 1 compares the scaled optimal allocation to the calibration allocation for several suboptimal costs. Only the achieved CV's for *var1*, *var3*, *var4* and *var9* are considered as these are the only variance constraints violated for the suboptimal costs considered here. As anticipated, there is a linear, proportional increase in the realized CV's for the scaled allocation. In general, under calibration allocation, tolerance is given only to the variance constraints placed on *var1* and *var4*. And, while the achieved CV's on these latter variables are relaxed more for the calibration allocation than the scaled optimal allocation, the suboptimal calibration allocations still satisfy the original objectives of the sample for *var3* and *var9*. Hence, when the optimal allocation is too costly, calibration allocation seems to have an advantage over the scaled optimal allocation.

Figure 1 – Tolerance/cost curves for suboptimal costs



4. SUMMARY AND FUTURE WORK

As expected under the propositions and supported by the example, we saw that: for the optimal cost, the calibration allocation is equivalent to the optimal allocation; and, for a suboptimal cost, calibration allocation is equivalent to the optimal allocation for the revised variance constraints. Furthermore, when the optimal cost is not acceptable, calibration allocation has advantage over the scaled optimal solution in that it allows differential tolerance to the variance constraints.

Using PROC NLP to obtain calibration allocations requires the hard-coding of much of the a priori information. In order to make a more general application, a different tool, such as PROC IML, might be examined. Other future work might include adapting the method to other sampling schemes such as multi-stage and two-phase designs.

It is conjectured that minimization of the following objective function which is a product of cost and the penalty function may obtain the optimal allocation as a calibration allocation where a constant a>1 (e.g. 10) is added to keep it bounded away from zero.

$$F = \left(\sum_h c_h n_h\right)\sum_j \ln\left(a + \exp\left[\lambda^{--1}\left(\sum_h n_h^{-1} x_{hj} - 1\right)\right]\right).$$

This function needs further investigation.

APPENDIX

Proof of Proposition 1: We consider two cases.

Case 1: Calibration allocation satisfies all variance constraints. Then, by the uniqueness of optimal allocation (Kokan and Khan 1966) the two must be equivalent.

Case 2: At least one variance constraint is not satisfied by calibration allocation. Then, for λ sufficiently small, $\sum_{h=1}^{H} n_h^{-1} x_{hj} - 1 > 0$ for at least one $j = 1,...,J$, and consequently $F \to \infty$. Under optimal allocation, all constraints are satisfied. Hence, $\sum_{h=1}^{H} n_h^{-1} x_{hj} - 1 \le 0$ for all $j = 1,...,J$ and consequently $F \le J \ln(2)$. However, by definition, the calibration allocation minimizes F for λ sufficiently small. Therefore, we have a contradiction. Hence the calibration allocation must satisfy all variance constraints, and it follows from Case 1 that the two must be equivalent.

Proof of Proposition 2: First we prove the lemma: let $\tilde{v}_j > v_j$ for at least one j=1,…,J be the realized variances for sufficiently small λ under the calibration allocation $\{\tilde{n}_{h,cal}^{(0)}\}$ with constraints $\hat{V}_j \le v_j$ and reduced cost $\tilde{\tau}$; let $\{\tilde{n}_{h,cal}^{(1)}\}$ be the calibration allocation for reduced cost $\tilde{\tau}$ with relaxed constraints $\hat{V}_j \le \max(v_j, \tilde{v}_j) = \tilde{v}_j^{(1)}$; and let $\{\tilde{n}_{h,cal}^{(2)}\}$ be the calibration allocation for reduced cost $\tilde{\tau}$ with revised constraints $\hat{V}_j \le \tilde{v}_j$. For λ sufficiently small, $\{\tilde{n}_{h,cal}^{(2)}\}$ satisfies all variance constraints with equality because there exists $\{\tilde{n}_{h,cal}^{(0)}\}$ with realized variances $\tilde{v}_j$. Consequently $F = J\ln(2)$. Next, while satisfying the reduced cost $\tilde{\tau}$, the realized variances cannot be less than $\tilde{v}_j$; otherwise, this would have been possible with $\{\tilde{n}_{h,cal}^{(0)}\}$ for smaller λ. Also, while satisfying $\hat{V}_j = \tilde{v}_j$, the calibration allocation cannot make cost less than $\tilde{\tau}$; otherwise, smaller variances should have been realized under $\{\tilde{n}_{h,cal}^{(0)}\}$. Thus, $\{\tilde{n}_{h,cal}^{(2)}\}$ has realized variances $\hat{V}_j = \tilde{v}_j$ for cost $\tilde{\tau}$.

Now consider the optimal cost τ_{rev}^{opt} for revised constraints $\hat{V}_j \le \tilde{v}_j$ and assume that $\tau_{rev}^{opt} < \tilde{\tau}$. So, by Proposition 1, there must also exist a calibration allocation with cost lower than $\tilde{\tau}$. However, this contradicts what was previously stated. Hence, $\tau_{rev}^{opt} = \tilde{\tau}$ and, by uniqueness of the optimal allocation, $\{\tilde{n}_{h,cal}^{(2)}\}$ must be equivalent to the optimal allocation. This implies that $\{\tilde{n}_{h,cal}^{(2)}\}$ is a unique calibration allocation for realizing $(\tilde{v}_j, \tilde{\tau}_j)$. This means that any other calibration allocation that realizes $(\tilde{v}_j, \tilde{\tau}_j)$ must be the same, i.e. $\{\tilde{n}_{h,cal}^{(0)}\} = \{\tilde{n}_{h,cal}^{(1)}\} = \{\tilde{n}_{h,cal}^{(2)}\}$.

Now to prove Proposition 2, suppose that the optimal cost under the relaxed constraints $\hat{V}_j \le \tilde{v}_j^{(1)}$ is less than the reduced cost $\tilde{\tau}$. Then from Proposition 1, it follows that the cost under the calibration allocation can be reduced further for the variance constraints $\hat{V}_j \le \tilde{v}_j^{(1)}$. This is a contradiction because it follows from the lemma that $\tilde{\tau}$ is the minimum cost that allows the variance constraints to be satisfied exactly at $\tilde{v}_j$. So, the optimal allocation with $\hat{V}_j \le \tilde{v}_j^{(1)}$ gives rise to $\tau_{rel}^{opt} = \tilde{\tau}$, which by Proposition 1implies that it is equivalent to the calibration allocation for $(\tilde{v}_j^{(1)}, \tilde{\tau}_j)$.

ACKNOWLEDGMENT

The first author's research was supported in part by a grant from the Natural Sciences and Engineering Research Council of Canada held at Carleton University under an Adjunct Research Professorship.

REFERENCES

Bethel, J. (1989). Sample allocation in multivariate surveys. *Survey Methodology*, **15**(1), 47-57.

Bethel, J.W. (1985). An optimum allocation algorithm for multivariate surveys. *American Statistical Association 1985 Proceedings of the Section on Survey Research Methods*, 209-212.

Chromy, J.R. (1987). Design optimization with multiple objectives. *American Statistical Association 1987 Proceedings of the Section on Survey Research Methods*, 194-199.

Cochran, W.G. (1977). *Sampling Techniques* (3rd ed). New York: John Wiley & Sons.

Estevao, V. (1998). Optimum allocation for one-stage stratified SRSWOR designs. Statistics Canada, internal report, October 1998.

Kokan, A.R. and Khan, S. (1963). Optimum allocation in multivariate surveys: an analytical solution. *Journal of the Royal Statistical Society B*., **29**, 115-125.

Rao, J.N.K and Singh, A.C. (1997). A ridge shrinkage method for range-restricted weight calibration in survey sampling. *American Statistical Association 1997 Proceedings of the Section on Survey Research Methods*, 57-65.