Statistics Canada Statistique Canada

Methodology Branch

Business Survey Methods Division

Direction de la méthodologie

Division des méthodes d'enquêtes-entreprises

Canadä

WORKING PAPER NO. BSMD-91-0011E
METHODOLOGY BRANCH

CAHIER DE TRAVAIL NO. BSMD-91-011E
DIRECTION DE LA MÉTHODOLOGIE

SAMPLE ALLOCATION FOR ESTIMATING MULTIPLE COMMODITY OUTPUTS OF MANUFACTURING INDUSTRIES

M.A. Rahim and S. Currie
Business Survey Methods Division
Statistics Canada

September 20, 1991

SAMPLE ALLOCATION FOR ESTIMATING MULTIPLE COMMODITY OUTPUTS OF MANUFACTURING INDUSTRIES

M.A. Rahim and S. Currie
Business Survey Methods Division

RÉSUMÉ

Le problème de l'estimation des valeurs de livraisons pour près de deux milles produits des industries manufacturières est abordé dans cet étude. Étant donné qu'un plan d'échantillonnage aléatoire stratifié sera utilisé pour l'enquête annuelle des établissements manufacturiers, une question fort importante qu'il faut se poser est celui-ci: Quelle est la meilleure façon de déterminer la taille de l'échantillon et de l'allouer à travers les strates? C'est un problème général d'optimisation de la taille et de l'allocation de l'échantillon dans le cas multivarié. Le problème est particulièrement difficile quand un très grand nombre de variables sont impliquées. Dans le cas où les bornes supérieures d'erreur d'échantillonnage des estimés sont fixées d'avance, une procédure itérative utilisant la programmation convexe peut-être utilisée. Cette procedure est passablement compliquée et donc recommandable seulement pour des problèmes dont le nombre de variables est "Modèrement grand" (Bethel, 1989). Cette procédure est néanmoins exposée pour le bénéfice du lecteur. Cependant, le but principal de l'étude réside plutôt dans la mise au point d'une méthode alternative, plus accessible et également satisfaisant pour tous les cas rencontrés en pratique. Cette alternative est illustrée grâce à l'utilisation des donnés de 1986 de l'industrie du vêtement. Même si la procedure ne garantit pas obligatoirement que les erreurs d'échantillonnage de tous les estimés individuels soient inférieures à la borne supérieure fixée, l'étude a démontré que la très grande majorité des estimés ont respecté la contrainte. Dans l'autre cas, c'est-à-dire celui où le coût de l'enquête est fixé d'avance, on a également développé une formule exacte d'allocation. Il a été démontré que cette formule est une extension de l'allocation de Neyman au cas multivarié et peut donc être utilisé dans le cas de l'estimation des valeurs de livraisons.

ABSTRACT

This study deals with the problem of estimating the values of shipments of nearly two thousand commodities of the manufacturing industries. Given that a stratified random sampling design will be used for the annual survey of the manufacturing establishments, an important question is how best to determine the sample size and allocate to the different strata. This is a general problem of optimizing sample size determination and allocation in the multivariate case, specially made difficult when a very large number of variables are involved. In the case when upper bounds to the sampling errors of the estimates are preassigned, an iterative procedure using convex programming can be used. This procedure is quite complicated and therefore suitable only for "moderately sized problems" (Bethel, 1989). Although we have explained this procedure for the general reader, the main purpose of our study was to investigate an alternative method, easy and satisfactory for all practical purposes. Such an alternative is presented here and illustrated using Statistics Canada's 1986 census data on clothing industry. Although the procedure is not expected to ensure that the sampling errors of all the individual estimates will remain below their preassigned upper bounds, this study has shown that a large majority of them did not exceed such upper bounds. In the other situation when cost of the survey is preassigned we have also derived a closed form allocation formula. It has been shown that this formula is an extension of Neyman Allocation in the multivariate case and can be used for the purpose of commodity output estimation.

SAMPLE ALLOCATION FOR ESTIMATING MULTIPLE COMMODITY OUTPUTS OF MANUFACTURING INDUSTRIES

M.A. Rahim and S. Currie
Business Survey Methods Division
Statistics Canada

CONTENTS

APPENDIX 1

SAMPLE ALLOCATION FOR ESTIMATING MULTIPLE COMMODITY OUTPUTS OF MANUFACTURING INDUSTRIES

M.A. Rahim and S. Currie
Business Survey Methods Division
Statistics Canada

1. INTRODUCTION

Under the Business Survey Redesign Project (BSRP) it is mandatory that all business surveys eventually use the Central Frame Data Base (CFDB) to extract their sampling frames and carry out functions related to the frame. For estimating commodity outputs of manufacturing industries, it is intended that all manufacturing establishments belonging to the integrated portion (IP) of the CFDB will be covered by a census and the establishments belonging to the non-integrated portion (NIP) will be covered by annual sample survey. For the latter the required data will be captured through an "other characteristics questionnaire" (OCQ) and the sampled units will be a subsample of the tax master sample. This is all the more necessary since basic financial data are not captured by the OCQ and therefore must be obtained from tax records. This subsample selection from the tax master sample will be based on a stratified random sampling design.

However, in this sampling procedure a difficulty arises because we are interested in estimating a very large number of commodity outputs and not just one. The question is how to determine the total sample size and an allocation rule to different strata such that either for a given cost, or for given upper bounds to the sampling errors we can obtain acceptable estimates of the total value of shipments for each of these commodities. It is this specific issue that has been addressed in this study.

In the case when the total cost of the survey is preassigned a closed form formula has been derived which is shown to be an extension of the Neyman allocation in the multivariate case. In the other situation when upper bounds to the sampling errors are preassigned, an iterative procedure using convex programming may be used. This procedure, however, is quite complicated and is therefore suitable for moderately sized problems only. We are confronted with the task of estimating about 2,000 commodity outputs (i.e. variables). Although we have explained and outlined the nature of the convex programming procedure for the general reader, the main purpose of our study was to investigate an alternative method which is easy and promising for estimating large number of commodities. Such an alternative is presented in this report and examined in the light of the 1986 census data relating to the clothing industry.

2. NATURE OF THE PROBLEM

In this section we will explain, in a rather simple manner, the underlying nature of the problem.

Suppose the population consists of N units, i.e. manufacturing establishments. Each unit manufactures one or more of the p distinct commodities represented by the variables x_1, $x_2,\ldots,x_p$. Total values of shipments of these commodities will be denoted by X_1, $X_2,\ldots,X_p$. Estimates of these totals based on stratified random sampling design will be denoted by $\hat{X}_1,\ \hat{X}_2,\ldots,\hat{X}_p$.

With a stratified random sampling design these estimates can be obtained under two different situations. First: total cost of the survey may be fixed. In that case we can determine the sample sizes in different strata in such a way that the sampling error of the estimates is minimum. Second: we decide to accept preassigned upper bounds to the sampling errors of the estimates. In that case we can determine the sample sizes in different strata in such a way that the cost is minimum and the sampling errors of the estimates do not exceed the preassigned upper bounds.

2.1 WHEN COST IS PREASSIGNED:

If the cost of surveying a unit is the same in each stratum, the total cost will depend on the total sample size n. In that case a preassigned cost implies a preassigned value of the total sample size n.

For such a preassigned value of n the best allocation of sample size n_{jh} to stratum h for estimating a specified commodity total X_j, is given by Neyman allocation,

$$n_{jh} = \frac{N_h\ S_{jh}}{\sum\limits_h N_h\ S_{jh}} * n \qquad (2.1.1)$$

where N_h is the total number of units and S_{jh} denotes the population standard deviation of the variable x_j in that stratum. However, this sample size n_{jh}, although best for the purpose of estimating X_j, will not be the best for estimating other commodity totals X_k; $k \neq j = 1,2,\ldots,p$; because of variation in S_{jh} for different j. Hence, the question arises, how to determine a single allocation of sample sizes $(n_1, n_2,\ldots,n_L)$, to L strata, such that estimates of all the commodity totals $(\hat{X}_1,\ \hat{X}_2,\ldots,\hat{X}_p)$ could be claimed as jointly "best", in some well-defined manner. The solution of this problem is given in section 3.

2.2 WHEN UPPER BOUNDS TO THE SAMPLING ERRORS ARE PREASSIGNED:

Alternatively, we may decide to accept preassigned upper bounds to the sampling error of the estimates and then look for the best allocation of sample sizes to strata by minimizing the cost. For a particular estimate of commodity total $\hat{X}_j$ the sampling error is usually expressed in terms of its coefficient of variation $cv\ (\hat{X}_j)$. For a given upper bound to this coefficient of variation $cv(\hat{X}_j) = \mu_o$ where μ_o may represent, say, 3%, 5%, 10%, etc., the best sample size n_h in h-th stratum is obtained by the formula.

$$n_{jh} = \frac{N_h\ S_{jh} \sum_h N_h\ S_{jh}}{(X_j\ .\ \mu_o)^2 + \sum_h N_h\ S_{jh}^2}\ ;\ h = 1,2,\ldots,L \qquad (2.2.1)$$

But then again, this best allocation for estimating X_j would not be the best for estimating other commodity totals X_k, $k \neq j = 1,2,\ldots,p$.

A point to be noted here is that we may decide to fix the same upper bound to the sampling error for estimating each of the p commodities. In other words, we may desire that $cv(\hat{X}_1) = cv(\hat{X}_2) = \ldots = cv(\hat{X}_p) = \mu_o$, or we may decide to fix different upper bounds to the sampling errors i.e. $cv(\hat{X}_1) \leq \mu_1\ ,\ cv(\hat{X}_2) \leq \mu_2\ ,\ \ldots, cv(\hat{X}_p) \leq \mu_p$.

In any case the same question arises, namely, how to determine a single allocation of sample sizes $(n_1, n_2, \ldots, n_L)$ such that the coefficients of variation of the estimates of the p commodity totals do not exceed their upper bounds while keeping the cost at minimum. A solution to this problem is also presented in section 3.

3. PROPOSED SOLUTION

In this section we will present a solution to the above problems that seems to be suitable, particularly in the context of manufacturing industries where we are confronted with the task of estimating values of shipments of about two thousand commodities. We will treat two different situations separately:

(i) cost is preassigned and we want to minimize the sampling errors of the estimates

(ii) upper bounds to the sampling errors of the estimates are preassigned and we want to minimize the cost

3.1 BEST ALLOCATION WHEN COST IS PREASSIGNED:

If we can assume that the cost of enumerating a unit is the same in each stratum, a preassigned cost is equivalent to a preassigned total sample size n, where $n = n_1 + n_2 + \ldots + n_L$. In the previous section we explained that given a preassigned n, there cannot be one single allocation $(n_1, n_2, \ldots, n_L)$ such that each individual $cv(\hat{X}_1)$, $cv(\hat{X}_2)$, ..., $cv(\hat{X}_p)$ is minimum. Given this fact, the only other option is to seek and settle for some compromise solution. One approach is to define a suitable function of $cv(\hat{X}_1)$, $cv(\hat{X}_2)$, ..., $cv(\hat{X}_p)$, which gives us a measure of joint variability of $\hat{X}_1$, $\hat{X}_2$, ..., $\hat{X}_p$. Let us consider such a function as

$$Y = W_1 cv^2(\hat{X}_1) + W_2 cv^2(\hat{X}_2) + \ldots + W_p cv^2(\hat{X}_p) \qquad (3.1.1)$$

where we assume that $\hat{X}_1$, $\hat{X}_2$, ..., $\hat{X}_p$, are independently distributed and W_j; $j = 1,2,\ldots,p$; are given weights reflecting the importance of the estimates $\hat{X}_1$, $\hat{X}_2$, ..., $\hat{X}_p$. Our problem would then reduce to minimizing Y subject to the condition that cost is fixed.

This formulation would have certain implications that should be clearly understood.

(i) Y is a measure of variability in an aggregate sense. In other words, we are assuming that when p is very large i.e. the number of commodity outputs to be estimated is very large, we would be more concerned with the joint variation of all the estimates rather than the variation of any individual estimate. Given this premise it makes sense to seek and settle for an allocation that will make Y minimum.

(ii) If, however, there is some variable x_k which has high variability in the population, then $CV(\hat{X}_k)$ will be large. Hence, $W_k CV^2(\hat{X}_k)$ will be a dominant term in Y. Minimization of Y and the resulting allocation n_1, $n_2, \ldots, n_L$, will then be largely

determined by minimization of $CV(\hat{X}_k)$. If x_k is a rare or unimportant variable, this will result in an allocation that optimizes the estimate of an unimportant variable at the expense of other important variables.

(iii) Nevertheless, this is analogous to an extreme value situation and can be countered by assigning appropriate weights W_j, j= 1, 2, ..., p. If x_k is that unimportant, one has the option of assigning a very small value to W_k. In the extreme case one can put $W_k = 0$, i.e. one can ignore the variable x_k in determining the allocation.

However, the overriding concern here is that we have to estimate as many as two thousand commodity outputs. Therefore, we decide to settle for a concept of "best" allocation in an aggregate sense. In other words, we define an allocaiton $n_1, n_2, \ldots, n_L$ to be the best if it will make Y minimum subject to the condition that cost is fixed.

For a stratified random sample,

$$cv^2\ (\hat{X}_j) = \left[\sum_h N_h\ (N_h - n_h)\ \frac{S_{jh}^2}{n_h}\right] \Big/ X_j^2\ ;\ j=1, 2, \ldots, p \tag{3.1.2}$$

Substituting these in 3.1.1 the full algebraical expression for Y can be written. The cost function is $C = \sum_j \sum_h C_{jh}\, n_h$ where C_{jh} represents unit cost of the j-th variable in the h-th stratum. Minimizing Y subject to the condition that C is fixed the solution for n_h is obtained as

$$n_h = \frac{\sqrt{\left\{\sum_j (W_j N_h^2 S_{jh}^2) \Big/ X_j^2\right\} \Big/ \left\{\sum_j C_{jh}\right\}}}{\sum_h \sqrt{\left\{\sum_j (W_j N_h^2 S_{jh}^2) \Big/ X_j^2\right\} \Big/ \left\{\sum_j C_{jh}\right\}}} * n;\ h= 1, 2, \ldots, L \tag{3.1.3}$$

If we are willing to assume $W_j = 1$, j=1,2,...,p, i.e. all the variables are of equal importance, then the above formula reduces to

$$n_h = \frac{\sqrt{\left\{\sum_j (N_h^2 S_{jh}^2) \Big/ X_j^2\right\} \Big/ \left\{\sum_j C_{jh}\right\}}}{\sum_h \sqrt{\left\{\sum_j (N_h^2 S_{jh}^2) \Big/ X_j^2\right\} \Big/ \left\{\sum_j C_{jh}\right\}}} * n;\ h = 1, 2, \ldots, L \qquad (3.1.4)$$

If the unit cost C_{jh} is constant, i.e. $C_{jh} = C$ then $\sum_{j=1}^{p} C_{jh} = \sum_{j=1}^{p} C = pC$. In that case the formula 3.1.4 further reduces to

$$n_h = \frac{\sqrt{\left\{\sum_j (N_h^2 S_{jh}^2) \Big/ X_j^2\right\}}}{\sum_h \sqrt{\left\{\sum_j (N_h^2 S_{jh}^2) \Big/ X_j^2\right\}}} * n;\ h = 1, 2, \ldots, L \qquad (3.1.5)$$

Full derivation of the formula, using two different methods, is given in appendix 1. It is also shown that this formula is an extension of Neyman allocation in the multivariate case.

3.1.1 CHOICE OF WEIGHTS:

It is evident from 3.1.3 that the allocation n_h in the h-th stratum is dependent on the assigned values of weights W_j, j= 1, 2,...,p. The question as to how best to assign these values is beyond the scope of this study. However, the following remarks could be helpful in understanding the nature of the problem.

(I) Values of W_j should be chosen in accordance with the relative importance of the estimators $\hat{X}_j$ - equivalently, relative importance of the variables x_j; j= 1, 2, ..., p. Note that under the situation when cost is preassigned we can choose any arbitrary set of values of W_j and obtain the optimum allocation n_h; h= 1, 2, ..., L; by minimizing the aggregate measure of variability Y. But that minimum value of Y may turn out to be large and may not be acceptable to us. In that case we can alter the values of W_j to bring the value of Y down to an acceptable level. The problem is, in the absence of any objective rule or criterion, those altered values have to be intuitive or judgemental.

(ii) In certain cases one might postulate that the more the variability of a variable x_j in the population, the less is its importance for the purpose of our estimation. In that case one may use the inverse of the estimate of the population variance as the weight i.e. $W_j = 1/\hat{V}(x_j)$; $j = 1, 2, . . . , p$.

(iii) In a situation like estimation of nearly 2000 commodity outputs, a common sense approach may be good enough for all practical purposes. For example, we can sort all the commodities into three groups such that the first group is of very little importance, the second group - a large majority - is of equal importance, while the third group - presumably a few - may be considered twice as important as the second group. In that case we can assign W= 0, 1, 2, to the commodities belonging to the first, second, and third group respectively.

For the purpose of our study where the central theme is to develop some method of allocation that is easy and good enough for all practical purposes, we assigned equal weights to all the commodities.

3.2 BEST ALLOCATION WHEN UPPER BOUNDS TO THE SAMPLING ERRORS ARE PREASSIGNED:

In this case we decide to preassign upper bounds to the sampling errors of the estimates $\hat{X}_j$; j= 1, 2,...,p; determined arbitrarily based on our perception of importance of the variables. In other words, we choose

$$cv\ (\hat{X}_j) \leq \mu_j ;\ j = 1, 2, \ldots, p \qquad (3.2.1)$$

where μ_j is any given value and then try to determine an allocation (n_1, n_2, n_3,...,n_L) such that the coefficients of variation of the estimates $\hat{X}_1, \hat{X}_2, \hat{X}_3, . . . , \hat{X}_p$, do not exceed the corresponding preassigned values $(\mu_1, \mu_2, . . . , \mu_p)$, respectively, subject to the condition that the total cost is minimum.

This is a complex mathematical problem and no exact theoretical solution has been obtained so far. However, a solution can be obtained resorting to iterative procedures using convex programming. A number of authors have dealt with this problem. Notable among the contributions are those of Dalenius (1957), Yates (1960), Kokan (1963), Hartley (1965), Kokan and Khan (1967), Chatterjee (1968, 1972), Huddleston, Claypool, and Hocking (1970), Chromy (1987), and Bethel (1985, 1989). Without going into the

mathematical details we will briefly explain here the motivation and theoretical basis of convex programming. We will also point out its limitations when we have to deal with large number of variables.

3.2.1 CONVEX PROGRAMMING AND ITS IMPLICATIONS:

In a recent paper, Bethel (1989) has given an improved version of the algorithm for convex programming. Following our notation let us write $\overline{X}_j$ as the estimate of a population mean obtained through stratified random sampling. Then it is known that

$$V(\overline{X}_j) = \sum_{h=1}^{L} \frac{N_h^2 S_{jh}^2}{N^2 n_h} - \sum_{h=1}^{L} \frac{N_h S_{jh}^2}{N^2} \tag{3.2.2}$$

where S_{jh}^2 is the population variance of the j-th variable x_j in the h-th stratum, $j=1, 2,\ldots,p$; $h=1, 2,\ldots,L$. N_h and n_h represents population and sample size respectively in the h-th stratum. $\sum\limits_h N_h = N$ and $\sum\limits_h n_h = n$ are the total population size and sample size respectively.

Taking the usual cost function $C = \sum\limits_h n_h C_h$ our problem is, we want to minimize

$$C = \sum_h n_h C_h \tag{3.2.3}$$

subject to the conditions

$$V(\overline{X}_j) = \sum_h \frac{N_h^2 S_{jh}^2}{N^2 n_h} - \sum_h \frac{N_h S_{jh}^2}{N^2} \le v_j;\ j=1, 2,\ldots,p \tag{3.2.4}$$

where v_j is some preassigned value depending on the importance of the variable $\mathbf{x}_j$. Writing

$y_h = \frac{1}{n_h}$; $a_{jh} = \frac{N_h^2 S_{jh}^2}{N^2 v_j}$; $k_j = 1 + \sum\limits_h \frac{N_h S_{jh}^2}{N^2 v_j}$; We can write 3.2.3 and 3.2.4 as

$$g(y) = \sum_h \frac{C_h}{y_h} \tag{3.2.5}$$

$$\sum_h a_{jh} y_h \le k_j;\ j = 1,\ 2,\ \ldots,\ p \tag{3.2.6}$$

Mathematically this means finding a minimum of the function g(y) subject to the conditions $\sum_h a_{jh} \overset{y_h}{\wedge} \le k_j;\ j = 1,\ 2,\ \ldots,\ p$.

Geometrically, one can visualize that a set $(y_1,\ y_2, \ldots, y_L)$ is merely a point in a L dimensional space. The conditions 3.2.6 demarcates a region R within that space. The convex programming works on an iterative procedure such that only the points in R are successively chosen converging to a unique set say $(y_1^0,\ y_2^0,\ \ldots,\ y_L^0)$ for which the cost function g(y) is minimum.

It is to be noted that Bethel has neglected the second term of the expression for $V(\bar{X}_j)$, presumably, for subsequent computational ease, and has used the approximate expression

$$V(\bar{X}_j) = \sum_h \frac{N_h^2 S_{jh}^2}{N^2 n_h} \tag{3.2.7}$$

In that case $k_j = 1$ and our problem reduces to finding a minimum of the function $g(y)$ subject to the conditions $\sum_h a_{jh}\, y_h \le 1$; j= 1,2,...,p. Equivalently, this means finding the minimum of the function F(y) where

$$F(y) = g(y) + \sum_j \lambda_j \left(\sum_h a_{jh}\, y_h - 1\right) \tag{3.2.8}$$

and λ_j ; j= 1, 2,...,p; are the Lagrange multipliers. Differentiating F(y) with respect to y_h ; h= 1, 2,...,L and λ_j ; j= 1, 2,...,p and equating to zero we get

$$
\begin{aligned}
\frac{\delta F(y)}{\delta y_h} &= -\frac{c_h}{y_h^2} + \sum_j \lambda_j a_{jh} = 0 \; ; \qquad h = 1,\, 2,\, \ldots, L \\
\frac{\delta F(y)}{\delta \lambda_j} &= \sum_h a_{jh} y_h - 1 = 0 \; ; \qquad j = 1,\, 2,\, \ldots, p
\end{aligned}
\qquad (3.2.9)
$$

Solving these equations, if one could get the values $y_1 = y_1^*\, ,\; y_2 = y_2^*\, ,\; \ldots, y_L = y_L^*$ then the set of values $\left(n_1 = \frac{1}{y_1^*}\, ,\; n_2 = \frac{1}{y_2^*}\, ,\; \ldots, n_L = \frac{1}{y_L^*}\right)$ would be the optimum allocations.

Unfortunately, an exact closed form solution of these equations is not possible. Alternatively, using Kuhn-Tucker theorem (1951) it is possible to write an expression for y_h^* as

$$
y_h^* = \frac{\sqrt{c_h}}{\sqrt{\left(\sum_{j=1}^{p} \alpha_j^* \, a_{jh}\right)} \; \sum_{k=1}^{L} \sqrt{\left\{c_k \sum_{j=1}^{p} \alpha_j^* a_{jh}\right\}}} \; ; \; h = 1,\, 2,\, \ldots, L \qquad (3.2.10)
$$

$$
if \; \sum_{j=1}^{p} \alpha_j^* a_{jh} > 0
$$

where $\alpha_j^* = \lambda_j / \sum_{j=1}^{p} \lambda_j$ (see Bethel (1989) for details). Note that this is still not a solution for the value of Y_h^* because α_j^* is unknown. However, this expression helps us to find out an approximate value for α_j^* through an iterative process. Bethel has given the details of the steps involved. An initial value $\alpha_j^{(1)}$; j= 1, 2,...,p; is chosen. Following those steps one can arrive at a value $\alpha_j^{(n)}$, at the n-th iteration, such that $|\alpha_j^{(n+1)} - \alpha_j^{(n)}| < \epsilon$; j= 1, 2,...,p, where ϵ is a preassigned convergence criterion. One can then substitute the value of $\alpha_j^* = \alpha_j^{(n)}$; j= 1, 2,...,p in 3.2.10 and get the optimum allocation $n_h = 1/y_h^*$; h = 1, 2,...,L.

Obviously, convex programming as outlined above, is theoretically attractive when we seek least cost allocation under the conditions that the sampling errors of the estimates of *each individual* variable do not exceed their preassigned upper bounds. However, it must be emphasized that it has certain difficulties when we have to deal with a very large number of variables as explained below.

(i) The number of iterations necessary depends on the number of strata (L) and the number of variables (p). The example used by Bethel for illustrative purposes consists of only 4 variables and 6 strata. He states "the algorithm converges quickly for most *moderately sized* problems" and the "run times vary *considerably* depending on the magnitude of the problem...". He also cautions that the "...labour involved in creating files and other preparatory tasks" is of much greater concern than merely the run times. In the context of manufacturing industries we are concerned with the estimation of about 2000 commodity outputs (i.e. variables). Obviously, in that case the convex programming would be too cumbersome if not altogether impossible.

(ii) Bethel also points out that "the convex programming approach gives the optimal solution to the defined problem but the resulting cost may not be acceptable so that a *further search* is *usually* required for an optimal solution...". This, however, can be done by "scaling down to the allowable budget directly and the effects of this on the precision of sample estimates can be directly determined". This means that in the case of large number of variables almost certainly we would run into a situation where we would have to increase the upper bounds of the allowable variance constraints for the estimates. In other words, the sampling errors of the estimates would have to exceed the levels that we had originally set.

(iii) Note that the value of v_j depends on the importance of the variable x_j - the more important it is, the smaller is the value of v_j we want to assign. The problem of assigning suitable values to v_j is similar to that of assigning values to W_j which we discussed in section 3.1.1. For any arbitrary set of values of $v_1, v_2, \ldots, v_p$ convex programming will lead to the optimum allocation n_h; h= 1, 2,...,L; by minimizing the cost C. But that minimum value of C may turn out to be large and may not be acceptable to us. In that case we can bring the value of C down to an acceptable level either by increasing all the values of v_j proportionately or by

altering them without maintaining the proportionality. Here again, the problem is, in the absence of any objective rule or criterion, those altered values have to be intuitive or judgemental.

Presumably, these as well as some other theoretical considerations led Bethel to state that "the problem of solving the convex optimization still remains".

This is why we believe that some other simpler approach, particularly for the purpose of our commodity output estimation, is worth investigating. One such approach will now be explained and illustrated in the subsequent sections of this report.

3.2.2 AN ALTERNATIVE APPROACH:

Let us assume that we agree on a preassigned coefficient of variation μ_0 (it may be 10%, 5%, 2%, etc.) such that the sampling errors of the estimates of the commodity totals $\hat{X}_1, \hat{X}_2, \ldots, \hat{X}_p$ do not exceed μ_0. In other words we want to ensure

$$cv(\hat{X}_j) \le \mu_0 \text{ ; for all } j = 1, 2, \ldots, p \qquad (3.2.11)$$

The corresponding optimum sample size for the single variable x_j in h-th stratum would be given by

$$n_{jh} = \frac{N_h S_{jh} \sum_h N_h S_{jh}}{(X_j \mu_0)^2 + \sum_h N_h S_{jh}^2} \qquad (3.2.12)$$

For the p variables $x_1, x_2, \ldots, x_p$ the optimum sample sizes in the h-th stratum would be $(n_{1h}, n_{2h}, \ldots, n_{ph})$. From this set of p sample sizes let us choose the maximum one and denote it by $n_{h(mx)}$. We then take $n_{h(mx)}$ as the sample size allocated in h-th stratum. We will refer to this allocation procedure as the "maximum rule". Obviously, under this allocation procedure all the p constrains at 3.2.11 would be satisfied. But the total sample size $n = \sum_h n_{h(mx)}$ would be large and therefore the cost would be high. Nevertheless, it will be worth investigating empirically if the cost is really too high to offset the simplicity of the procedure.

Instead of choosing the maximum, we could have chosen, say, the median of the p sample sizes $(n_{1h}, n_{2h},\ldots,n_{ph})$ and take $n_{h(med)}$ as the sample size allocated to h-th stratum, h= 1, 2,...,L. We refer to this allocation procedure as the "median rule". In fact, one could choose any percentile value $n_{h(p)}$ and take the same approach, namely, choose the p-th percentile of the distribution of sample sizes in h-th stratum.

Under any of these allocation procedures it would be worth investigating how far the constraints at 3.2.11 are violated and to what extent the cost is affected.

These investigations have been carried out based on 1986 clothing industry data. The results are presented in section 4.

4. AN EMPIRICAL STUDY

In this section we will present the results of an empirical study relating to the effects of the sample allocation procedures, explained earlier.

For an initial investigation, the 1986 data on the clothing industry (SIC 24) were used. Under this industry sector there are 15 commodity groups that come under import control. The miscellaneous class, namely, "Other Controlled Commodities", was left out of the study. The remaining 14 commodity groups, included in this study are shown in the table below.

Table 4.1

COMMODITY GROUPS INCLUDED IN THE STUDY

IMPORT CONTROL GROUPS	CLOTHING COMMODITY
32	Winter Outerwear
37	Pants, shorts and overalls
39	Blouses, shirts, t-shirts and sweatshirts
40	Sleepwear, bathrobes and dressing gowns

IMPORT CONTROL GROUPS	CLOTHING COMMODITY
44	Swimwear
45	Underwear
46	Outer jackets, coats and shopcoats
47	Sportscoats, blazers and fine suits

Table 4.1 (continued)

41	Rainwear
42	Dresses, skirts, sets and co-ordinates
43	Foundation garments

48	Leather coats and jackets
49	Tailored collar shirts (Men's and Boy's)
50	Sweaters, pull-overs and cardigans

4.1 RESULTS WITH PREASSIGNED UPPER BOUNDS TO THE SAMPLING ERRORS:

A total of 2256 manufacturing establishments i.e. units were covered in this study. For each unit, total value of shipments of one or more of the 14 commodity groups - represented as the variables x_1, x_2,...,x_{14} - were considered as our observations. A combination of Province and revenue class was defined as a stratum. Table 4.1.1 below shows population size N_h, and the allocated sample size n_h in h-th stratum under the maximum, median, and 75th percentile rule. These allocations were obtained with an upper bound to the CVs equal to 10% and following the sample selection procedure, earlier described, under section 3.2.2.

For those strata with population size equal to or less than 5, all units were included in the sample. The largest revenue class in each province was defined as a take-all stratum for which also all units were included in the sample.

Table 4.1.1

SAMPLE SIZE UNDER DIFFERENT ALLOCATION PROCEDURE

STRATA		N_h	TARGET CV 10%		
PROV	REVENUE CLASS		MAX.	75th Pct	MED
NFLD	25,000-99,999	1	1	1	1
PEI	100,000-249,999	1	1	1	1
NS	25,000-99,999	1	1	1	1
	100,000-249,999	1	1	1	1
	250,000-749,999	2	2	2	2
	750,000-3,099,999	3	3	3	3

Table 4.1.1 (continued)

	3,100,000 and over	2	2	2	2
NB	100,000-249,999	2	2	2	2
	250,000-749,999	3	3	3	3
	750,000-1,699,999	3	3	3	3
	1,700,000 and over	1	1	1	1
PQ	25,000-99,999	129	4	3	3
	100,000-249,999	243	12	5	5
	250,000-749,999	357	33	11	9
	750,000-3,899,999	479	332	119	67
	3,900,000 and over	285	285	285	285
ONT	25,000-99,999	52	3	2	2
	100,000-249,999	54	5	2	2
	250,000-749,999	116	16	6	3
	750,000-6,999,999	268	254	78	45
	7,000,000 and over	56	56	58	56
MAN	25,000-99,999	5	5	5	5
	100,000-249,999	4	4	4	4
	250,000-749,999	20	3	2	1
	750,000-4,099,999	33	30	13	2
	4,100,000 and over	20	20	20	20
SASK	25,000-99,999	1	1	1	1
	100,000-249,999	1	1	1	1
	750,000 and over	4	4	4	4
ALTA	25,000-99,999	8	1	1	1
	100,000-249,999	5	5	5	5
	250,000-749,999	11	4	3	1
	750,000-3,499,999	12	7	5	1
	3,500,000 and over	4	4	4	4
BC	25,000-99,999	10	1	1	1
	100,000-249,999	9	1	1	1
	250,000-749,999	16	2	2	2
	750,000-1,999,999	18	8	3	2
	2,000,000 and over	18	18	18	18
	TOTAL	2256	1139	680	570

Note:all fractions have been rounded up to the next higher integer.

We now have three different sample allocations corresponding to three different rules. For each allocation we can now compute the values of the coefficient of variations of the estimates of the commodity totals $\hat{X}_1, \hat{X}_2, \ldots, \hat{X}_{14}$, using the formula 3.1.2. Table 4.1.2 below shows how much these coefficients of variations differ from one another due to three different allocation procedures.

Table 4.1.2

COEFFICIENT OF VARIATION OF ESTIMATES OF COMMODITY TOTALS UNDER DIFFERENT SAMPLE ALLOCATIONS

Estimates of commodity totals $\hat{X}_i$	CV($\hat{X}$) under different sample allocation procedure		
	Maximum (Target CV = 10%)	75th Percentile (Target CV = 10%)	Median (Target CV = 10%)
$\hat{X}_1$	0.038	0.103	0.193
$\hat{X}_2$	0.009	0.030	0.047
$\hat{X}_3$	0.020	0.060	0.085
$\hat{X}_4$	0.035	0.093	0.131
$\hat{X}_5$	0.025	0.068	0.150
$\hat{X}_6$	0.015	0.046	0.065
$\hat{X}_7$	0.020	0.051	0.069
$\hat{X}_8$	0.053	0.239	0.338
$\hat{X}_9$	0.030	0.073	0.101
$\hat{X}_{10}$	0.026	0.079	0.125
$\hat{X}_{11}$	0.027	0.062	0.086
$\hat{X}_{12}$	0.061	0.151	0.257
$\hat{X}_{13}$	0.029	0.089	0.124
$\hat{X}_{14}$	0.031	0.091	0.145
AVERAGE CV	0.030	0.088	0.137

It is now easy to draw several conclusions from tables 4.1.1 and 4.1.2.

FIRST: The maximum rule does ensure - as expected - that the CV of estimates of each commodity total are less than the preassigned value. For example, when the preassigned value was 10% the CVs, on an average, were as low as 3%. However, the total sample size had to be as high as 1139 for a population of size 2256. In other words, the cost becomes prohibitive and such a sample selection procedure would not be desirable. Clearly, we would reach the same conclusion if we chose any other preassigned values of CV such as 5% or 2%, etc.

SECOND: On the other hand, for a preassigned CV of 10%, if we had used the median rule, the total sample size would be reduced to 570 for a population of size 2256, but the CV's of the estimates of commodity totals, on an average, would be as high as 14%. In other words, the cost would be quite acceptable but the sampling error would far exceed the preassigned value. It is also seen that the coefficients of variation of the estimated shipment totals of 9 commodities - out of 14 - exceed the preassigned value of 10%. Thus, adopting the median rule does not seem to be a desirable procedure either.

THIRD: The above findings lead, naturally, to the next question, namely, what would happen if we selected sample size based on 75th percentile? Intuitively, one can think that this sample selection procedure might lead to a total sample size not too large (i.e. cost would not be prohibitive) and at the same time the CVs, on an average, would not be far off from the preassigned value. This is confirmed by the figures in table 4.1.1 and 4.1.2. Under this selection procedure total sample size comes out to be 680, which is not too large, and the average CV is 9%, close to the target CV of 10%. For two of the estimates, however $(\hat{X}_8,\ \hat{X}_{12})$, the CVs are much larger than 10%.

It, therefore, appears that a sample selection procedure based on 75th percentile rule will most likely provide a solution that is simple, and acceptable for all practical purpose. This procedure will be particularly suitable when we have to deal with a large number of variables as in the case of estimation of the commodity totals in manufacturing industries. (NOTE: The formula 3.1.5 requires computation of the values of S_{jh}^2, the stratum variances in the population. For the year 1986 we had the population data. For subsequent years - in the absence of population data - we would have to estimate the stratum variances by using the commodity output values from samples).

4.2 RESULTS WITH PREASSIGNED COST

In many practical situations, as in the case of annual survey of manufacturing industries, the number of units to be surveyed is determined in advance based on cost considerations. In that case the formula 3.1.5 is appropriate. This formula, as we have shown (see appendix), is based on an extension of Neyman allocation to the multivariate case. Using this formula we computed the sample sizes in different strata for a given total sample size n=680, obtained earlier by using the 75th percentile rule. In other words, we preassigned the same cost as we would have if we used the 75th percentile rule. Tables 4.2.1 and 4.2.2 below shows the comparative results with regard to the allocation of sample sizes and the values of the coefficients of variation.

Table 4.2.1
COMPARATIVE RESULTS: ALLOCATION OF SAMPLE SIZES

STRATA		POPULATION SIZE	SAMPLE SIZE (75th Pct.rule)	SAMPLE SIZE (Extended Neyman allocation formula)
PROV	REVENUE CLASS	N_h	n_h	n_h
NFLD	25,000-99,999	1	1	1
PEI	100,000-249,999	1	1	1
NS	25,000-99,999	1	1	1
	100,000-249,999	1	1	1
	250,000-749,999	2	2	2
	750,000-3,099,999	3	3	3
	3,100,000 and over	2	2	2
NB	100,000-249,999	2	2	2
	250,000-749,999	3	3	3
	750,000-1,699,999	3	3	3
	1,700,000 and over	1	1	1
PQ	25,000-99,999	129	3	2
	100,00-249,999	243	5	5
	250,000-749,999	357	11	13
	750,000-3,899,999	479	119	114

Table 4.2.1 (continued)

	3,900,000 and over	285	285	285
ONT	25,000-99,999	52	2	2
	100,000-249,999	54	2	2
	250,000-749,999	116	6	6
	750,000-6,999,999	266	78	90
	7,000,000 and over	56	56	56
MAN	25,000-99,999	5	5	5
	100,000-249,999	4	4	4
	250,000-749,999	20	2	2
	750,000-4,099,999	33	13	11
	4,100,000 and over	20	20	20
SASK	25,000-99,999	1	1	1
	100,000-249,999	1	1	1
	750,000 and over	4	4	4
ALTA	25,000-99,999	8	1	1
	100,000-249,999	5	5	5
	250,000-749,000	11	3	2
	750,000-3,499,999	12	5	3
	3,500,000 and over	4	4	4
BC	25,000-99,999	10	1	1
	100,000-249,999	9	1	1
	250,000-749,999	16	2	2
	750,000-1,999,999	18	3	3
	2,000,000 and over	18	18	15
	TOTAL	2256	680	680

TABLE 4.2.2
COMPARATIVE COEFFICIENTS OF VARIATION OF THE ESTIMATES OF COMMODITY TOTALS

COMMODITY TOTALS	CV OF THE ESTIMATES ((Under allocation based on the 75th Pct. rule)	CV OF THE ESTIMATES (Under extended Neyman allocation)
X_1	0.103	0.103
X_2	0.030	0.029
X_3	0.060	0.058
X_4	0.093	0.093
X_5	0.068	0.079
X_6	0.046	0.044
X_7	0.051	0.048
X_8	0.239	0.201
X_9	0.073	0.072
X_{10}	0.079	0.075
X_{11}	0.062	0.060
X_{12}	0.151	0.152
X_{13}	0.089	0.084
X_{14}	0.091	0.089
AVERAGE CV	0.088	0.085

An examination of the Tables (4.2.1) along with (4.2.2) reveals the following interesting fact. We had seen that if sampling error is preassigned, a sample allocation procedure based on 75th percentile rule may lead to the determination of sample sizes in each stratum such that, on an average, the sampling errors of multiple variables do not exceed a preassigned value. However, we did not know if the cost incurred by that process - determined by the total sample size n= 680 - was too high and unreasonable.

It has been found that by using an extension of the Neyman allocation formula for the same cost, the minimum sampling error attainable - on an average - is $CV_{Min} = 0.085$ which is almost identical with what we had obtained ($CV_{75} = 0.088$) by using the 75th percentile rule and that the CV for each estimate is either same or slightly smaller except for X_5. In other words, this investigation has shown that a sample allocation procedure based on the 75th percentile rule may give us, on an average, sampling errors of estimates in the multiple variables situation which do not exceed a preassigned value. And at the same time the cost incurred by the process is close to what we could have if we used the extended Neyman allocation formula.

5. CONCLUSION AND SUMMARY

This study is concerned with the problem of estimation of the total values of shipments of specified commodities or commodity groups based on annual survey of manufacturing industries. Given that a stratified simple random sampling design will be adopted for selecting the manufacturing establishments (i.e. units) from the tax master sample, one important question is how to determine the total sample size and allocate it to the different strata defined by a combination of Provinces and revenue classes.

We have pointed out that this is the well known general problem of optimizing sample size determination and allocation in the multivariate case. We have explained the exact nature of the difficulties that arise in the multivariate situation either under the constraint of preassigned upper bounds to the sampling errors of the estimates or under the constraint of preassigned cost. The problem, the related concepts, and the solutions proposed, have all been presented in a way expected to be helpful for survey planning purposes.

In the case of preassigned upper bounds to the sampling errors, a number of computer algorithms were suggested as far back as in 1967 by Kokan and Khan (14) and recently in 1989 by Bethel (1). These algorithms are all based on a complicated, iterative process known as convex programming. We have explained the underlying concept as well as the mathematical reasonings related to such programming, particularly with reference to Bethel's recent work. This procedure, although theoretically attractive, appears to be suitable, according to Bethel "for moderately sized problems" only.

The main purpose of this study was to investigate an alternative method which is easy and promising for the purpose of estimating a very large number - maybe well over 2000 - of commodity outputs. Such an alternative is presented in this report. The applicability of this procedure has been examined in the light of actual Statistics Canada data on clothing industry sector (1986). What we have found is that a sample allocation procedure based on "75th percentile rule" works quite well from the point of view of simplicity and cost effectiveness. The procedure is flexible in the sense that it can be altered to any other "p-th percentile rule" to reduce the desired number of variables for which sampling errors of estimates exceed the preassigned upper bounds. It is also flexible in the sense that one can preassign different upper bounds to the sampling errors of the estimates depending on their importance (we used a target CV of 10% for all the variables for the sake of simplicity). It must be understood, however, that the procedure seems to be suitable based on this empirical study. It does not ensure, theoretically, that the sampling error of each individual estimate will not exceed its preassigned upper bound, although this study has shown that a large majority of them did not exceed. In fact only 2 out of 14 sampling errors of the estimates had significantly exceeded their preassigned upper bounds.

Quite often, as in the case of annual survey of manufacturing industries by Statistics Canada, we preassign the survey cost and therefore the total number of units that we want to survey is predetermined. In that case, for the purpose of estimating multiple commodity outputs, we have developed a closed form formula for allocation of sample sizes to different strata. We have shown that this formula is an extension of Neyman allocation in the multivariate case. This extended formula ensures a minimum value of an aggregate measure of the variability of all the estimates. We, therefore, suggest that this method of sample allocation can be conveniently used for the purpose of commodity output estimation.

ACKNOWLEDGEMENT

The authors acknowledge the comments of the referees and the Director of the Business Survey Methods Division that led to substantial improvement of this report. Danielle Lalande and Wisner Jocelyn rendered considerable help in programming and computation. Thanks are also due to Linda Lafontaine who took special care in typing this report.

APPENDIX 1

DERIVATION OF FORMULA 3.1.5 (Method 1)

For a stratified design let $\hat{X}_j$ denote the estimate of the total value X_j of a variable x_j; j = 1, 2,...,p. It is well known (see Cochran 3rd edition. p-93) that

$$CV(\hat{X}_j) = \frac{1}{X_j}\sqrt{\left\{\sum_h^L N_h(N_h-n_h)\frac{S_{jh}^2}{n_h}\right\}} \quad . \tag{1}$$

where N_h, n_h, and S_{jh}^2 are the population size, sample size, and variance of x_j in the h-th stratum; h = 1, 2, ...,L. We define a function Y of $CV(\hat{X}_j)$; j = 1, 2,...,p, as

$$Y = \sum_j W_j CV^2(\hat{X}_j) \tag{2}$$

we also write the usual cost function as

$$C = \sum_j\sum_h C_{jh}n_h \tag{3}$$

where C_{jh} is the per unit cost for the j-th variable in the h-th stratum. Our problem is to minimize Y with respect to $n_1, n_2,...,n_L$, subject to the condition that C is fixed. Equivalently, we have to minimize a function F where we write

$$\begin{aligned} F &= \sum_j W_j\, CV^2\, (\hat{X}_j) + \lambda\, (\sum_j\sum_h C_{jh}\, n_h - C) \\ &= \sum_j \frac{W_j}{X_j^2} \sum_h N_h\, (N_h-n_h)\, \frac{S_{jh}^2}{n_h} + \lambda\, (\sum_j\sum_h C_{jh}\, n_h - C) \\ &= \sum_j\sum_h \frac{W_j\, N_h^2\, S_{jh}^2}{X_j^2 n_h} - \sum_j\sum_h \frac{W_j N_h\, S_{jh}^2}{X_j^2} + \lambda\, (\sum_j\sum_h C_{jh}\, n_h - C)\, , \end{aligned} \tag{4}$$

where λ is the lagrange multiplier. Differentiating (4) with respect to n_h; h= 1, 2,...,L, and putting it equal to zero we can write

$$-\sum_j \frac{W_j\ N_h^2\ S_{jh}^2}{X_j^2\ n_h^2} + \lambda\ \left(\sum_j C_{jh}\right) = 0$$

$$\text{or } n_h\sqrt{\lambda} = \sqrt{\left[\left\{\sum_j (W_j\ N_h^2\ S_{jh}^2)/X_j^2\right\}/\left\{\sum_j C_{jh}\right\}\right]} \tag{5}$$

Now summing over all h we can write

$$n\sqrt{\lambda} = \sum_h \sqrt{\left[\left\{\sum_j (W_j\ N_h^2\ S_{jh}^2)/X_j^2\right\}/\left\{\sum_j C_{jh}\right\}\right]} \tag{6}$$

Then, dividing (5) by (6) we get

$$\frac{n_h}{n} = \frac{\sqrt{\left[\left\{\sum_j (W_j\ N_h^2\ S_{jh}^2)/X_j^2\right\}/\left\{\sum_j C_{jh}\right\}\right]}}{\sum_h \sqrt{\left[\left\{\sum_j (W_j\ N_h^2\ S_{jh}^2)/X_j^2\right\}/\left\{\sum_j C_{jh}\right\}\right]}}$$

or

$$n_h = \frac{\sqrt{\left[\left\{\sum_j (W_j\ N_h^2\ S_{jh}^2)/X_j^2\right\}/\left\{\sum_j C_{jh}\right\}\right]}}{\sum_h \sqrt{\left[\left\{\sum_j (W_j\ N_h^2\ S_{jh}^2)/X_j^2\right\}/\left\{\sum_j C_{jh}\right\}\right]}} * n \tag{7}$$

This n_h, j= 1, 2,...,L, is the optimum allocation in h-th stratum for a fixed cost C.

SPECIAL CASE 1

If we assume W_j = 1; j= 1, 2,...,p, i.e. weights are the same, then the formula (7) reduces to

$$n_h = \frac{\sqrt{\left[\left\{\sum_j (N_h^2\ S_{jh}^2)/X_j^2\right\}/\left\{\sum_j C_{jh}\right\}\right]}}{\sum_h \sqrt{\left[\left\{\sum_j (N_h^2\ S_{jh}^2)/X_j^2\right\}/\left\{\sum_j C_{jh}\right\}\right]}} * n \tag{8}$$

Usually, we also assume that the unit cost C_{jh} is constant i.e. C_{jh} = c. Then $\sum_{j=1}^{p} C_{jh} = \sum_{j=1}^{p} c$ = pc. Hence, the formula further reduces to

$$n_h = \frac{\sqrt{\left\{\sum_j (N_h^2 S_{jh}^2) / X_j^2\right\}}}{\sum_h \sqrt{\left\{\sum_j (N_h^2 S_{jh}^2) / X_j^2\right\}}} * n \tag{9}$$

This completes the derivation of our formulae 3.1.4 and 3.1.5.

SPECIAL CASE 2

In the case of a single variable i.e. j= 1 we can drop the subscript j and the formula (9) reduces to

$$n_h = \frac{\sqrt{(N_h^2 S_h^2)/X^2}}{\sum_h \sqrt{(N_h^2 S_h^2)/X^2}} * n$$

$$\text{or } n_h = \frac{N_h S_h}{\sum N_h S_h} * n; \; h=1,\, 2,\, \ldots, L \tag{10}$$

which is the well known Neyman allocation formula. Our formula 3.1.5 is thus an extension of the Neyman allocation in the multivariate case.

DERIVATION OF FORMULA 3.1.5 (METHOD 2):

Our objective is to minimize $Y = \sum_j W_j CV^2(\hat{X}_j)$ subject to the condition that the cost $C = \sum_h C_h n_h$ is fixed. Writing the expression for $CV(\hat{X}_j)$ and assuming the weights to be equal i.e. W_j = 1, we get

$$Y=\sum_j \frac{1}{X_j^2}\sum_h N_h(N_h-n_h)\frac{S_{jh}^2}{n_h}$$

$$=\sum_j\sum_h \frac{N_h^2 S_{jh}^2}{n_h X_j^2}-\sum_j\sum_h \frac{N_h S_{jh}^2}{X_j^2}$$

$$=\sum_h \frac{N_h^2}{n_h}\sum_j \frac{S_{jh}^2}{X_j^2}-\sum_h N_h\sum_j \frac{S_{jh}^2}{X_j^2}$$

$$= Y_1-Y_2 \ (suppose) \quad ,$$

where S_{jh}^2 denotes the population variance of the variable x_j in the h-th stratum. (11)

Note that Y_2 is independent of n_h; h = 1, 2,...,L. Thus, minimizing Y is equivalent to minimizing Y_1 with respect to n_1, n_2,...,n_L.

Now suppose, the cost C is increased by a small amount δ. Consequently, we may increase the sample size n_r in the r-th stratum to $n_r+\frac{\delta}{C_r}$; C_r being the per unit cost of enumeration in the r-th stratum. Then the value of Y_1 would decrease due to the reduction in the r-th term, equal to K, where

$$\frac{N_r^2}{n_r}\sum_j \frac{S_{jr}^2}{X_j^2}-\frac{N_r^2}{n_r+\delta/C_r}\sum_j \frac{S_{jr}^2}{X_j^2} = K \qquad (12)$$

In order that the allocation n_1, n_2,...,n_L is optimum this amount K must remain same i.e. constant irrespective of the stratum r that we might choose for increasing the sample size. If this were not the case, the allocation $n_1, \ n_2, \ldots, n_L$ could be improved by shifting sample units between strata without increasing the overall cost C.

From (12), for any stratum h, we can therefore write

$$N_h^2 \sum_j \frac{S_{jh}^2}{X_j^2}\left[\frac{1}{n_h}-\frac{1}{n_h+\delta/C_h}\right]=K; \quad h=1,\ 2,\ \ldots,L$$

$$\text{or} \quad N_h^2 \sum_j \frac{S_{jh}^2}{X_j^2}\left[\frac{\delta}{n_h(n_hC_h+\delta)}\right]=K$$

$$\text{or}\ \delta N_h^2 \sum_j \frac{S_{jh}^2}{X_j^2}=Kn_h(n_hC_h+\delta) \tag{13}$$

Assuming δ to be small with respect to n_hC_h we can write (13) as

$$\delta N_h^2 \sum_j \frac{S_{jh}^2}{X_j^2}=Kn_h^2C_h$$

$$\text{or} \quad n_h=\sqrt{\frac{\delta}{K}}\sqrt{\frac{N_h^2}{C_h}\sum_j \frac{S_{jh}^2}{X_j^2}} \tag{14}$$

Summing over h we get

$$\sum_h n_h=n=\sqrt{\frac{\delta}{K}}\sum_h\sqrt{\frac{N_h^2}{C_h}\sum_j \frac{S_{jh}^2}{X_j^2}} \tag{15}$$

Dividing (14) by (15) we get

$$\frac{n_h}{n}=\frac{\sqrt{\frac{N_h^2}{C_h}\sum_j \frac{S_{jh}^2}{X_j^2}}}{\sum_h\sqrt{\frac{N_h^2}{C_h}\sum_j \frac{S_{jh}^2}{X_j^2}}}$$

$$or \quad n_h = \frac{\sqrt{\frac{N_h^2}{C_h}\sum \frac{S_{jh}^2}{X_j^2}}}{\sum_h \sqrt{\frac{N_h^2}{C_h}\sum_j \frac{S_{jh}^2}{X_j^2}}} * n \qquad (16)$$

If C_h is constant in every stratum i.e. $C_h = C$, then (16) reduces to .

$$n_h = \frac{\sqrt{\sum_j (N_h^2 S_{jh}^2) / X_j^2}}{\sum_h \sqrt{\sum_j N_h^2 S_{jh}^2) / X_j^2}} * n; \quad h = 1, 2, \ldots, L \qquad (17)$$

which is the same as the formula 3.1.5.

REFERENCES

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

2. BETHEL, J.W. (1985). *An optimum allocation algorithm for multivariate surveys. Proceedings of the Survey research Section, American Statistical Association, 209-212.*

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

4. CHATTERJEE, S. (1968). *Multivariate stratified surveys. Journal of the American Statistical Association, 63, 530-534.*

5. CHATTERJEE, S. (1972). *A study of optimum allocation in multivariate stratified surveys. Skandinavisk Actuarietidskrift, 55, 73-80.*

6. COCHRAN, W.G. (1953). *Sampling Techniques. New York: Wiley.*

7. DALENIUS, T. (1953). *The multivariate sampling problem. Skandinavisk Actuarietidskrift, 36, 92-102.*

8. 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.*

9. FORSYTH, G.E. (1968). *On the asymptotic directions of the s-dimensional optimum gradient method. Numerische Mathematik, 11, 57-76.*

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

11. HUDDLESTON, H.F., CLAYPOOL, P.L., and HOCKING, R.R. (1970). *Optimum sample allocation to strata using convex programming. Applied Statistics, 19, 273-278.*

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

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

14. 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.*

15. KUHN, H.W. and TUCKER, A.W. (1951). *Nonlinear programming. Proceedings 2nd Berkeley Symposium Mathematical Statistics and Probability.*

16. LUENBERGER, D.G. (1984). *Linear and Nonlinear Programming. Reading, Massachusets: Addison-Wesley.*

17. NEYMAN, J. (1934). *On the two different aspects of the representative method: the method of representative sampling and the method of purposive sampling. Journal of the Royal Statistical Society, 558-625.*

18. YATES, F. (1960). *Sampling Methods for Censuses and Surveys. London: Charles Griffin and Company.*