Statistics Canada Statistique Canada

Methodology Branch

Business Survey Methods Division

Direction de la méthodologie

Division des méthodes d'enquêtes-entreprises

Canada

WORKING PAPER NO. BSMD-93-006E

CAHIER DE TRAVAIL NO. BSMD-93-006E

METHODOLOGY BRANCH

DIRECTION DE LA MÉTHODOLOGIE

SOME APPLICATIONS OF LINEAR PROGRAMMING TECHNIQUES TO SAMPLING

by

Ioana Schiopu-Kratina

May 1993

SOME APPLICATIONS OF LINEAR PROGRAMMING TECHNIQUES TO SAMPLING

Ioana Schiopu-Kratina

Résumé

Ce rapport traite, à la manière de Des Raj (1956) et à l'aide de la programmation linéaire, le problème de maximisation du chevauchement de deux échantillions tirés successivement d'une population ainsi que le problème de l'échantillonnages contrôlé. Le rapport présente une approche générale et analyse l'existance des solutions. L'approche couvre des cas classiques de la théorie d'échantillonnage (Keyfitz, Lahiri, Goodman et Kish, Cochran) ainsi que des développements plus recents (J.N.K. Rao et Nigam). On examine l'application possible du problème du transport dont la fonction coût est de type Monge, ainsi que l'utilisation de l'algorithme "Northwest Corner Rule", qui correspond à la méthode d'échantillonnage de Lahiri. On généralise des résultats connus au cas de trois unitées sélectionées d'une population stratifiée (trois strates).

Abstract

This technical report attempts to treat in a unified way the problem of maximizing the overlap of two samples drawn for two surveys and the problem of controlled sampling, using a linear programming approach, In the manner of Des Raj (1956). The general problem is formulated and the existence of the solution is discussed. It is shown that this general formulation covers classical sampling problems (Keyfitz, Lahiri, Goodman and Kish, Cochran), as well as recent developments (J.N.K. Rao and Nigam). The applicability of the transportation problem when the cost function is submodular is studied. It is shown that Goodman and Kish's example as well as Cochran's hospitals example can be solved using the Northwest Corner Rule algorithm, which corresponds to Lahiri's sampling scheme. A generalization to selecting 3 units from 3 strata is also given.

1. Introduction

The purpose of this technical report is to generalize and unify some of the recent results in the area of applications of linear programming techniques to maximizing the overlap of two samples drawn for two surveys and to controlled sampling.

Maximum overlap of two samples drawn for two different surveys is a classical sampling problem which appears in the literature in one form or another since 1951 (see Keyfitz (1951), Lahiri (1954)). It is a problem of great practical importance for any statistical agency that carries out repeated surveys. The idea is to draw the sample for the second survey so as to respect some requirements of that survey, (e.g. first order inclusion probabilities) while retaining as many units as possible from the sample selected in the first survey. Such procedures reduce to a minimum the cost of contacting new respondents and minimize the break in the time series of estimates. A variation of this idea is encountered in multi-purpose surveys in which units are selected for measuring, for example, two characteristics. In this instance, it is desirable to measure the two characteristics on units that are close in some sense, say geographically. Ideally, the same unit could be used for measuring both characteristics. This would greatly reduce the travelling and administrative costs as well as response-related problems.

Problems of controlled selection are treated in an article by Goodman and Kish (1950) in the context of stratified sampling. A stratification of the population into 2 strata is given and a sample of size 2 is drawn, with one unit selected from each stratum. A further restriction is imposed on the set of possible samples. Certain samples are declared preferred over others. In Goodman and Kish's example (1950), units of different types (costal, inland) form the preferred samples, presumably to obtain a better representation of the population and reduce the variance of the estimators. The population is not stratified according to types and it is by assigning larger probabilities of selection to preferred samples that the "controls" are exercised, while the first order inclusion probabilities are specified.

Of course, the probabilities of selection that give an optimal solution do not have to coincide with those corresponding to stratified sampling, where the selection is done independently in each stratum.

We are concerned here with the mathematical treatement of such sampling problems using linear programming techniques.

The problem of maximum overlap received a unified linear programming treatement in the monograph by Arthanari and Dodge (1981) (see problem 5.8.1). Formulation (3.3) of this report covers, in fact, most of the examples presented by Arthanari and Dodge.

Since the linear programming approach of Des Raj (1956), the problem of controlled selection has not received much attention until recently. Rao and Nigam (1990-91) present a general situation where linear programming techniques can be applied for controlled selection. Although the cost function has only values 0 or 1, the set-up does not have the simple structure of the classical cases.

The purpose of this work is firstly to give a general linear programming formulation that includes as particular models the maximum overlap of samples and controlled selection (see (3.3) and (3.3)' of this report). In terms of a simple model, we could have the following situation. A population is sampled on two different occasions. The first order inclusion probabilities are given on both occasions as well as the list of preferred (and therefore nonpreferred) samples. We would like to maximize the probability of selecting a preferred sample on the first occasion and retain it for the second occasion.

Secondly, we examine the existence and properties of the solution of the general problem stated above. In the course of this analysis, we show that in some instances the problem reduces to a transportation problem with a submodular cost matrix (see (4.2) of section 4) for which the Northwest Corner Rule (NWCR) algorithm gives an optimal solution. This is a greedy-type algorithm (Hoffman (1985)) that is very efficient. We apply it to the controlled sample example of Cochran (1977) p.126, which is of the same type as Goodman and Kish's example (1950). We obtain the same solution as Cochran, who used Lahiri's selection algorithm. Thus, Cochran's solution is optimal. This should come as no surprise, since the NWCR and Lahiri's selection scheme are identical algorithms so either can be used to find optimal solutions for transportation problems with submodular cost matrices. We indicate how the method could be generalized to selecting three units from a population divided into three strata, when three types of units are present (end of section 4).

The general formulation does not always lead to a transportation problem, even in the simple case of one survey. In fact, we are dealing with a transportation problem with additional constraints. Even obtaining a basic feasible solution does not follow the classical procedure of assigning maximum possible mass on the diagonal, as Example 5.1 illustrates. This example could, in fact, be generalized to produce a feasible solution under more general conditions then those of Proposition 3.1.

This paper is organized as follows. Section 2 presents a brief study of the literature. Section 3 contains the formulation of the general problem as well as the existence of the solution and examples. We show then that this formulation covers all the classical cases presented in Arthanari and Dodge (1981) as well as the controlled sampling formulated by Rao and Nigam (1990). In section 4 we present the applicability of the Northwest Corner Rule (NWCR) to transportation problems in general and to some instances of optimal controlled sampling in particular. Section 5 is devoted to the study of the structure of the optimal solution in the case of one survey. Further research could pursue the question of optimality of the procedures applied (see the proof's of Lemmas 5.1, 5.2) and their presentation in the form of an algorithm.

2. Brief study of literature.

This section contains a brief study of literature pertinent to the problem presented in this article. It is, by no means, an exhaustive study. The work presented in the articles by Des Raj (1956), Mitra and Pathak(1984), Causey et al (1985) and Rao and Nigam (1990) is outlined in this section.

The use of linear programming techniques in sampling appears in an article by Des Raj (1956). In his article, the problems of Lahiri (1954), Keyfitz (1951) and Goodman and Kish (1950) are set-up as linear programming problems and solved.

The problem of Lahiri arose in the context of a multipurpose survey in which villages were the sample units for measuring the land utilization area and the population. The probabilities of selection of a village, for the two characteristics to be measured, are, in general, different. In order to reduce the cost of the survey, it is desired that the villages selected for measuring these characteristics be "almost identical" in a geographical sense, so as to reduce the expense of, among other things, travelling.

Des Raj provides a solution using the simplex method due to Dantzig. He uses the matrix of real transportation costs between villages. In Lahiri's solution the villages are ordered in a serpentine manner. Lahiri's solution is an optimal solution for a transportation problem in which the cost function associated with a pair of villages (v_i, v_j) is $c(i,j) = |i-j|, i,j > 0$. Because of the serpentine labelling of the villages, this "cost" is often proportional to the real distance between villages. Lahiri's solution for the common probabilities of selecting the villages v_i and v_j in the surveys, $\{q(i,j)\}_{i,j=1},\ldots,N$, is very close to Raj's solution. This is a particular instance of the more general problem formulated in section 3 of this article (see Example 3.3). In the context of this article, we are dealing with the case of only one unit that has to be selected on each occasion and the probabilities of selecting the units on each occasion are given: p_i for the first survey, p'_i for the second, for unit u_i, $i = 1,\ldots,N$ (see also Example 3.2). One wishes to minimize the total cost $\sum_{j,i=1,\ldots,N} c_{ij}\, q(i,j)$, where $c_{ij} = |i-j|, i,j = 1,\ldots,N$. This formulation of Lahiri's problem appears in Arthanari and Dodge (1981) (problem 5.4.2). It is shown there that an optimal solution to this problem is obtained using the Northwest Corner Rule (NWCR) (Athanari and Dodge p. 248 and Results 5.6.1, 5.6.2).

In the above paper by Des Raj, the "preferred sample" problem of Goodman and Kish is also presented as a linear programming problem and solved using the simplex algorithm. We show that this problem, as well as the example that appears in Cochran (1977) p. 126 can be solved using the NWCR, because they can be set-up as transportation problems with the appropriate type of cost matrix. The advantage of doing so is the efficiency of this algorithm.

A paper by Mitra and Pathak (1984) presents a different approach to the integration of surveys. Algorithms for integrating two and three surveys are given. An integration of k-surveys is viewed as a cartesian product S^k with a probability $\mathcal{P}$, where S is the population and $\mathcal{P}$ is a probability with given marginals $\mathcal{P}_i$, $= 1,\ldots,k$. Only samples of size 1 are considered for each survey, so each integrated sample $x = (x_1,\ldots,x_k)$ consists of k-units.

Let $\nu(x)$ denote the number of distinct units in the sample x. The article gives algorithms for obtaining optimal integrated surveys with given marginals for which $E[\nu(X)]$ is a minimum, when $k = 2$ and $k = 3$. For $k = 2$, the algorithm assigns maximum possible mass (from the marginals) to samples in the integrated survey that contain the same unit twice, and calculates the remaining probabilities. It appears that, with the problem set-up as a transportation problem the algorithm is equivalent to that of finding a basic feasible solution or an optimum solution to Keyfitz's scheme with p_i , p_j' , $i,j = 1,\ldots,N$ as marginals: (see Problem 5.4.1 of Arthanri and Dodge (1981)). The algorithm for $k = 3$ is more complex. It is not clear that a transportation problem type approach is applicable, although there are similarities with the case $k = 2$ presented above.

The importance of using algorithms related to the transportation theory is emphasized by Causey, Cox and Ernst (1985). They use transportation theory to controlled selection problem, controlled rounding and maximizing the overlap between surveys. We give below a succinct description of their article.

The controlled rounding as a transportation problem was first presented by Cox and Ernst (1982). It consists of the replacement of a real number x by an adjacent integer $R(x)$ (base 1 rounding) such that tabular arrays remain tabular (see (1.2), (2.3) of their paper). More precisely, starting with a two-way table satisfying some additivity constraints, one obtains, by rounding, a new table which still satisfies the additivity constraints. Cox and Ernst (1982) also obtained optimal solutions (in the sense of minimizing an ℓ_p distance) subject to the restriction that integer values round to themselves. Controlled rounding is used in controlling statistical disclosure in tables of frequency counts and to prevent statistical disclosure in microdata release. It can also be applied to raking in two-way tables of counts.

The paper by Causey, Cox and Ernst uses a controlled rounding algorithm, based on a transportation problem for solving a controlled selection problem. The setting is different from that of Des Raj's paper (1956) or of the present paper. It can be described as follows:

A population is partitioned by two criteria of classification represented by a two-way table of $m \times n$ classification cells S. A sample of size n must be selected with s_{ij} the expected number of units in each cell. The purpose of the algorithm is to obtain n_{ij}, the actual number of sample units in the cell (i,j) and to control its deviation from s_{ij} , $1 \leq i \leq m+1$, $1 \leq j \leq n+1$. This deviation is at most equal to 1. The requirements of probability sampling are strictly maintained. In the initial two-way table of classification S, the entries are sums of probabilities. By controlled rounding, a finite sequence of arrays N_k and associated probabilities p_k are produced, where each N_k is a controlled rounding of S. In each cell (i,j) we obtain a random variable n_{ij} taking the values n_{ijk} with probabilities p_k , $k = 1,\ldots,\ell$, and $E[n_{ij}] = s_{ij}$, $1 \leq i \leq m+1$, $1 \leq j \leq n+1$.

As an example of the applicability of this procedure, the authors cite a redesign of the household survey conducted by the U.S. Bureau of the Census. Each row represents a stratum and each column represents a state (not a classification variable for the original stratification). One primary sampling unit (P.S.U.) must be selected from each stratum so the total of each row is 1. The entry in each cell of S is the sum of probabilities of the P.S.U.'s to be selected from that state. A controlled selection would produce N_k , a $\{0,1\}$ array with a single 1 in each row indicating, for each stratum, the state in which the P.S.U. would be located. From this state then, a single P.S.U. would be selected with probability proportional to size.

In this approach, controlled rounding is used to control the support of the sample, while maintaining the requirement of probability sampling. In our approach, the solution consists of probabilities which determine the sample design.

Causey et al. also discuss methods for maximizing the overlap between surveys. This is a classical problem of great practical interest for any statistical agency. Keyfitz (1951) presented an optimum procedure for selecting one P.S.U. per stratum when the strata are identical in the surveys. The case when the strata change from one survey to the next was studied by Perkins (1970) and Kish and Scott (1971). Fellegi (1966) considered the case of two P.S.U.'s per stratum selected without replacement with P.P.S. These procedures, however, are not optimal in the set-up of Causey et al. (or Arhanari and Dodge (1981)).

We use now the notation of Causey et al. in presenting their treatment of the maximum overlap between surveys as a transportation problem. Let x_{ij} represent the joint probabilities of selecting the sample I_i in the first survey, and S_j in the second survey $i = 1,\dots,m$, $j = 1,\dots,n$.

One wishes to maximize $\sum_{i=1}^{m} \sum_{j=1}^{n} c_{ij}\, x_{ij}$, subject to $x_{ij} \geq 0$, $\sum_{j=1}^{n} x_{ij} = p_i$, $\sum_{i=1}^{m} x_{ij} = \pi_j$, where π_j is the probability of selecting S_j , p_i is the probability of selecting I_i and c_{ij} is the number of P.S.U.'s in $I_i \cap S_j$, $i = 1,\dots,m$, $j = 1,\dots,n$. This formulation is similar to the more general Problem 5.8.1 of Arthanari and Dodge, with the obvious change in notation. However, the particular problem faced by Causey et al. is more complex. They are dealing with a multistage stratified design, with no restriction on the sample size. In the original design, the selection need not be performed independently in each stratum and so the values of p_i , $i = 1,\dots,m$ might not be easy to calculate from the selection probabilities of the original sampling plan.

One of the most recent articles on the use of linear programming in sampling is that of Rao and Nigam (1990). They obtain optimal solutions to controlled sampling problems using linear programming methods, as an alternative to the more complicated experimental design methods.

For a given population, the list of all possible samples S is given, of which S_1 represents the subset of nonpreferred samples. The authors discuss two types of optimal plans. Firstly, they consider plans with specified second order inclusion probabilities π_{ij} , $i,j = 1,\ldots,N$. The optimal controlled design $p_c(s)$, is a solution to the linear programming problem: minimize $\sum\limits_{s \in S_1} p(s)$, subject to: $p(s) \geq 0$, $s \in S$

and $\sum\limits_{i,j \in s} p(s) = \pi_{ij}$ $(i < j, i, j = 1,\ldots,N)$.

Secondly, they consider different constraints on $p(s)$, namely: $\sum\limits_{i \in s} p(s) = np_i$, $i = 1,\ldots,N$ and

$c(np_i)(np_j) \leq \sum\limits_{i,j \in s} p(s) \leq (np_i)(np_j)$ where np_i are first order inclusion probabilities

$i,j = 1,\ldots,N$, $0 < c < 1$. They correspond to optimal inclusion probability proportional to size plans for which there are stable, nonnegative, unbiased variance estimators. The applicability of these linear programming methods is illustrated throughout their paper with examples.

Last but not least, one must mention the monograph Mathematical Programming in Statistics, by Arthanari and Dodge (1981), that presents in a rigorous and general setting the background as well as the up-to-that date research in the field. Their results are cited throughout this article.

3. Mathematical formulation of the problem. Existence of the solution. Examples.

Consider a finite population P which is sampled on two different occasions. Let S be the set of all possible subsets of P. To distinguish the two occasions, we use prime for the corresponding mathematical symbols on the second occasion. We look at $P \times P'$ and all its possible subsets $S \times S'$. On this product space, we consider a probability function q , $0 \leq q(s,s') \leq 1$ for any $(s,s') \in S \times S'$. We must have:

(3.0)
$$\sum_{(s,s') \in S \times S'} q(s,s') = 1.$$

We may look at the set of pairs of samples corresponding to two given sampling plans as a subset of $S \times S'$. Then $q(s,s') = 0$ if s or s' is not a possible sample. The probability q induces marginal probabilities on each of the components (occasions). These are:

(3.1)
$$p(s) = \sum_{s' \in S'} q(s,s') \text{ , } p'(s') = \sum_{s \in S} q(s,s'), \text{ any } (s,s') \in S \times S'.$$

Assume that on each occasion we have $\pi_u (\pi'_{u'})$, the probability of selecting unit $u\,(u')$ on the first (second) occasion is given, for any $u \in P$ $(u' \in P')$. This means that p and p' must satisfy (3.2) for all $u \in P$, $u' \in P'$:

(3.2)
$$\sum_{s:u \in s} p(s) = \pi_u \text{ and } \sum_{s':u' \in s'} p(s') = \pi'_{u'}.$$

Putting together (3.1) and (3.2), we obtain the system of constraints:

(3.3a)
$$\begin{cases} \sum\limits_{\substack{s \in S \\ u \in s}} \sum\limits_{s' \in S'} q(s,s') = \pi_u & u \in P \\[2ex] \sum\limits_{\substack{s' \in S' \\ u' \in s'}} \sum\limits_{s \in S} q(s,s') = \pi'_u & u' \in P' \end{cases}.$$

Since the right hand side is a probability, (3.3a) implies that each $q(s,s') \leq 1$, $(s,s') \in S \times S'$. However, if we wish to recover $q(s,s')$ as a probability function from (3.3a) we must also impose, in addition, the conditions:

(3.3b)
$$0 \leq q(s,s'), \text{ for all } (s,s') \in S \times S'.$$

Furthermore, q must satisfy (3.0). This, however, follows from (3.3a) due to the properties of the inclusion probabilities.

We would want now the selection on the two occasions to be performed in an optimum way. This, of course, may exclude independent selections on each occasion.

For instance, a preferred sample is selected on the first occasion, we would like to retain it on the second occasion and preserve the inclusion probabilities on each occasion. Or we may want to maximize the overlap of units selected on two different occasions. Thus, a pair of samples (s,s') in which s' differs from s by at most one unit is preferred over a pair (s,s') with no units in common. These situations and many others could be modeled by introducing a cost function defined on $S\times S'$, $c(s,s')$, $(s,s')\in S\times S'$.

We then seek to minimize the total cost, i.e.

(3.3c)
$$\min_q \sum_{(s,s')\in S\times S'} c(s,s')\cdot q(s,s').$$

Alternatively, we may attach weights to each pair (s,s'), $0\leq w(s,s')\leq 1$ and find a maximum:

(3.3c')
$$\max_q \sum_{(s,s')\in S\times S'} w(s,s')\cdot q(s,s').$$

We refer henceforth to conditions (3.0) and (3.3a)-(3.3c) as condition (3.3) and to conditions (3.0) and (3.3a)-(3.3c') as (3.3'). Either constitutes a mathematical formulation of the problem discussed earlier.

Remark 3.1: Equation (3.3a) could be modified if the sampling plans on the two occasions are given in terms of second order inclusion probabilities, say π_{uv}, $u,v,\in P$ for the first survey. Then (3.3a) becomes:

(3.4)
$$\sum_{\{s\in S:(u,v)\in s\}} \sum_{s'\in S'} q(s,s') = \pi_{uv} \quad \text{all } u,v\in P.$$

The same modification could be made on the second equation (3.3a). Of course, in this instance the possible samples must contain more than two units.

Example 3.1: Assume that we are sampling on one occasion only and would like to match second order inclusion probabilities π_{uv}, $u,v \in P$. We have a set of nonpreferred samples $S_1 \subset S$ and give a cost of 0 to all preferred samples and a maximum cost of 1 to all $s \in S_1$. Then (3.4) becomes, if we take $S' = P'$ and $p(s) = q(s,P')$, all $s \in S$:

$$\sum_{\substack{s \in S \\ u,v \in s}} p(s) = \pi_{uv} \quad u,v \in P.$$

We also require $p(s) \geq 0$ and seek a minimum (in p) of the objective function $\varphi(p) = \sum_{s \in S_1} p(s)$.

But this is exactly the formulation (3.3) of Rao and Nigam (1990) with a slight change in notation. In addition, we also required (3.0), i.e. $\sum_{s \in S} p(s) = 1$. This can be shown to hold automatically.

Example 3.2: Assume that p and p' are known (see (3.1)) on both occasions and that there is no requirement for matching inclusion probabilities on either occasion. Since S, S' are finite sets (of cardinality N, say), we can number the samples so that each (s,s') corresponds to some (i,j), $i,j \leq N$. We have from (3.1) and (3.3) with $q_{ij} = p_{ij}$ the statement:

$$(3.5) \qquad \left\{ \begin{array}{l} \textit{minimize } \sum_i \sum_j c_{ij}\, p_{ij} \textit{, subject to:} \\ \sum_{j=1}^{N} p_{ij} = p_i \\ \sum_{i=1}^{N} p_{ij} = p_j' \qquad i,j = 1,\ldots,N \\ 0 \leq p_{ij} , \end{array} \right.$$

This is problem 5.8.1 of Arthanari and Dodge (1981). Notice that, since p and p' are probability functions, $\sum_{i=1}^{N} \sum_{j=1}^{N} p_{ij} = 1$ and (3.0) is therefore obvious. If, in this example we take $cij = \begin{cases} 0 \text{ if } i=j & i,j=1,...N \\ 1 \text{ if } i\neq j \end{cases}$ we obtain the mathematical formulation of Keyfitz's problem (1951) (see Artharari and Dodge, Problem 5.4.1).

Example 3.3: This is a particular case of Example 3.2, when all samples consists of one unit each. It could also be obtained directly from the general formulation (3.3) when each s (s') is a unit u (u'). In this case, $\{\pi_u\}_{u\in P}$ and $\{\pi'_{u'}\}_{u'\in P'}$ are probability functions.

Proposition 3.1: For prespecified first or second order inclusion probabilities, corresponding to existing sampling plans, an optimum solution to problem (3.3) (or (3.3')) always exists.

Proof: All we need to show is that the feasible region determined by (3.3) (or (3.3')) is nonempty. The hypotheses say that, for given π_u , $\pi'_{u'}$, $(u,u')\in P\times P'$ one can find probability functions p on S and p' on S' such that (3.2) holds, with π_u , $\pi'_{u'}$ pre-assigned. Once p and p' satisfy (3.2), (see Example 3.1) we must find q that satisfies (3.0) and (3.1). Note that, since p and p' are probabilities, if (3.1) holds then so does (3.0). Thus we must show that (3.1) has a solution in q, for given p , p'. But this is a classical transportation problem which is known to have a solution, i.e. the feasible region is nonempty. The optimum for any linear objective function is obtained at a vertex of the feasible region.

Remark 3.2: From the proof of the proposition, it is clear that the population does not have to be the same on both occasions. A general set-up would be the same with the only difference that the indexing sets are different in s and s'. However, the transportation problem does not require that the indexing be the same (see (1.1) of Causey, Cox and Ernest (1985)). Thus, the problem is relevant to selections from two different populations with an appropriately defined cost function.

Remark 3.3: The proof of the existence of the solution was done in two stages, following the set-up at the beginning of the section. This need not lead to the optimum solution, which is found by solving directly the system (3.3) or (3.3').

Remark 3.4: When the sample sizes are larger than one, and the order in which a sample is drawn in each survey is irrelevant, the matrix of coefficients of q in (3.3) is not a classical transportation matrix. The column corresponding to $q(s,s')$ in the matrix of constraints has $1's$ in more than two rows, as the following example shows.

Example 3.4: We have the same population of three units $P = \{u_1, u_2, u_3\}$ sampled on two occasions without replacement. The sample size on each occasion is $n = 2$, so there are three possible samples on each occasion. We list them: $s_1 = \{1,2\}$, $s_2 = \{1,3\}$, $s_3 = \{2,3\}$. The probability of selecting samples_1 in the first survey and s_2 in the second is q_{12}, q_{32} stands for the probability of selecting sample s_3 on the first occasion and s_2 on the second and so on. Equation (3.3a) corresponding to π_1 is:

$$q_{11} + q_{12} + q_{13} + q_{21} + q_{22} + q_{23} = \pi_1.$$

For π_2, we have:

$$q_{11} + q_{12} + q_{13} + q_{31} + q_{32} + q_{33} = \pi_2$$

and for π_1':

$$q_{11} + q_{21} + q_{31} + q_{12} + q_{22} + q_{32} = \pi_1'.$$

Therefore, the coefficient of q_{11} is equal to 1 in more than two equations. The entire matrix of coefficients in (3.3a) is:

$$\begin{array}{ccccccccc|cl}
q_{11} & q_{12} & q_{13} & q_{21} & q_{22} & q_{23} & q_{31} & q_{32} & q_{33} & & \\
1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & \pi_1 & \text{survey I} \\
1 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & \pi_2 & \\
0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & \pi_3 & \\
\hline
1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & \pi'_1 & \text{survey II} \\
1 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & \pi'_2 & \\
0 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & \pi'_3 &
\end{array}$$

__Example 3.5__: If we consider only one survey, as in Example 3.4, then matching the first order inclusion probabilities in the previous example reduces to three equations and the number of variables reduces to three as well: $q(s_1) = q_1$, $q_2 = q(s_2)$ and $q_3 = q(s_3)$ (see Example 3.1):

$$\begin{aligned} q_1 + q_2 &= \pi_1 \\ q_1 + q_3 &= \pi_2 \\ q_2 + q_3 &= \pi_3 \quad . \end{aligned}$$

The matrix of coefficients is:

$$\begin{array}{ccc|c}
q_1 & q_2 & q_3 & \\
\hline
1 & 1 & 0 & \pi_1 \\
1 & 0 & 1 & \pi_2 \\
0 & 1 & 1 & \pi_3
\end{array}$$

For a larger population, the number of possible samples of size $n < N$ is quite large and the number of columns in the matrix of coefficients is larger than the number of rows. For instance, if $N = 4$ and $n = 2$ we have 6 possible samples, thus six columns and four rows in the matrix of coefficients.

__Example 3.6__: Consider now the same population of example 3.4 and the possible samples $S = \{s_1 , s_2 , s_3\} = S'$.

As in example 3.2, we sample on two different occasions and the probabilities of selecting each sample on either occasion is known. With the notation of Example 3.2, we have:

$$\sum_{j=1}^{3} p_{ij} = p_i \ , \ i = 1,2,3 \ \text{and} \ \sum_{j=1}^{3} p_{ij} = p'_j \ , \ j = 1,2,3.$$

The matrix of coefficients in this case is a classical transportation matrix:

$$\begin{array}{ccccccccccl}
p_{11} & p_{12} & p_{13} & p_{21} & p_{22} & p_{23} & p_{31} & p_{32} & p_{33} & & \\
1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & p_1 & \text{survey I} \\
0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & p_2 & \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & p_3 & \\
\hdashline
1 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & p'_1 & \text{survey II} \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & p'_2 & \\
0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & p'_3 &
\end{array}$$

Example 3.6': In example 3.5, let us consider all possible samples. We do not identify the samples (i,j) and (j,i) $i \neq j$, $i,j = 1,2,3$. This corresponds to the situation when the order of drawing the sample is relevant. There are then 9 variables and the coefficient matrix is the same as in Example 3.6. The identification of q_{ij} and q_{ji} , $i \neq j$ and elimination of the columns q_{ii}, gives, in the upper part of the matrix in Example 3.6, the coefficient matrix of Example 3.5. The bottom half now repeats the same information as the top half and can therefore be ignored.

Remark 3.5: Example 3.5 could be viewed as the transportation problem 3.6-3.6' with the additional constraints: the samples (i,i) are excluded ($i = 1,2,3$, sampling without replacement) and $q_{ij} = q_{ji}$, $i \neq j$, $i,j = 1,2,3$.

Example 3.7: We are going to formulate in terms of a transportation problem the preferred sample example given in Cochran (1977) p. 126. Two strata are considered: large and small hospitals. Two types of ownership are of interest. One hospital has to be selected from each stratum and representation of each type of ownership is desired. This problem is similar to the problem of Goodman and Kish which was solved using the simplex algorithm by Des Raj (1957). We will show that, in fact, it can be solved using the NWCR (see section 4) which is a very efficient algorithm.

Firstly, the units in each stratum are reordered. Cardinals without prime represent one type of ownership, those with prime another. The revised order is presented below:

Large Hospitals	Small Hospitals
1	3'
2	4'
3'	5'
4'	1
	2

The revised order is essential for the algorithm described in Cochran, based on the selection of random numbers. It somehow resembles Lahiri's serpentine method as the reordering makes the desired pairs of units "close" in some distance.

We relabel the units of the two strata so that $\pi_i(s)$ is the inclusion probability in stratified sampling that corresponds to unit $u_i(s)$ of the small hospitals with the revised order, $i = 1,\ldots,5$. For example $\pi_1(s)$ is the inclusion probability corresponding to unit 3' of the small hospitals. Similarly, $\pi_4(\ell)$ is the inclusion probability corresponding to unit 4' of the large hospitals and so on. Because of the nature of stratified sampling, $\sum_{i=1}^{5} q(u_i(s)\ ,\ u_j(\ell)) = \pi_j(\ell)\ ,\ j = 1,\ldots,4$ and $\sum_{j=1}^{4} q(u_i(s)\ ,\ u_j(\ell)) = \pi_i(s)\ ,\ i = 1,\ldots,5$. One can see then that the matrix of coefficients is a transportation matrix.

4. The applicability of the Northwest Corner Rule (NWCR).

We start this section by discussing the extent of applicability of the NWCR in finding a basic feasible solution that is optimal for a transportation problem. For simplicity, we are selecting samples of size 1 from two different populations of size N and M. Let c,q: $N\times M\to R$, which could be viewed as vectors in R^{N+M} and define $c\cdot q$ (scalar product) as $\sum_{\substack{j\in N\\ j\in M}} c(i,j)\, q(i,j)$. Consider again the transportation problem: find q: $N\times M\to R_+$ such that:

$$\text{(4.1)}\qquad \begin{aligned} &\sum_{j\in M} q(i,j) = p(i)\,, \qquad i\in N \\ &\sum_{i\in N} q(i,j) = p'(j) \qquad j\in M \end{aligned}$$

where $\sum_{i\in N} p(i) = \sum_{j\in M} p'(j) = 1$ and p: $N\to R_+$, p': $M\to R_+$ are given. We intend to minimize $c\cdot q$, for c given, where q runs over all basic feasible solutions of (4.1).

Recall that a cost matrix $c(i,j)_{(i,j)\in N\times M}$ is submodular (or Monge) if, for all (i,j) , $1\le i\le N$, $1\le j\le M$, $(i+\ell,j+k)\in N\times M\,, \ell>0\,, k>0:,$

$$\text{(4.2)}\qquad c(i,j) + c(i+\ell\,, j+k) \le c(i,j+k) + c(i+\ell,j)$$

(see Ross (1983), p. 6).
We have:

Lemma 4.1: Assume that c is a submodular matrix and that q is a basic feasible solution of (4.1). Assume that, for some (i,j) , $1\le i\le N$, $1\le j\le M$, there exists $\ell,k\ge 1$ such that $q(i+\ell,j) > 0$, $q(i,j+k) > 0$. There exists then a basic feasible solution q^* of (4.1) such that $c\cdot q^* \le c\cdot q$. Furthermore, $q^*(i,j) > q(i,j)$.

Proof: The proof is very similar to the proof of Result 5.6.1 of Arthanari and Dodge (1981) and will not be given here.

The following result is well known. (see Hoffman, 1985)

Proposition 4.1: A basic feasible solution obtained by NWCR is optimal if the cost matrix is submodular.

Proof: The proof uses Lemma 4.1 and is identical to the proof of Result 5.6.2 of Arthanari and Dodge (1981).

Next, a criterion for submodularity is given.

Proposition 4.2: The function c: $N\times M\to R$ is submodular if and only if, for any $i,j \in N\times M$,

$$c(i+1\ ,\ j+1)\ -\ c(i+1\ ,\ j) \le c(i,j+1)\ -\ c(i,j) \tag{4.3}$$

Proof: Clearly, submodularity implies (4.3). Conversely, assume that $c(m+1\ ,\ n+1)\ -\ c(m+1\ ,\ n) \le c(m,n+1)\ -\ c(m,n)$, for all m,n. Let us now sum over m, as m runs between i and $i+\ell-1$. As the sum telescopes, we obtain:

$$c(i+\ell\ ,\ n+1)\ -\ c(i+\ell\ ,\ n)\ -\ c(i\ ,\ n+1)\ +\ c(i,n) \le 0.$$

Now, let us sum over n, as n runs from j to $j+k-1$:

$$\sum_{n=j}^{j+k-1} [c(i+\ell\ ,\ n+1)\ -\ c(i+\ell\ ,\ n)\ -\ c(i\ ,\ n+1)\ +\ c(i\ ,\ n)] =$$

$$c(i+\ell\ ,\ j+k)\ -\ c(i+\ell\ ,\ j)\ -\ c(i\ ,\ j+k)\ +\ c(i\ ,\ j) \le 0\quad .$$

This is precisely the definition of submodularity (4.2).

Remark 4.1: Proposition 4.2 is the discrete analog of Proposition 4.2 of Ross (1983). If $c(x,y)$ is the extension of c to R^2, then, if the second order derivatives exists, $\dfrac{\partial^2 c}{\partial x \partial y} \le 0$ is a criterion for submodularity.

Example 4.1: The function $c(i,j) = -i\cdot j$ is submodular.

Indeed, if we consider $c(x,y) = -xy$, $\dfrac{\partial^2 c(x,y)}{\partial x \partial y} = -1 < 0$. One cannot readily apply this procedure to $c(i,j) = |i-j|$ as $c(x,y) = |x-y|$ has no derivative if $x = y$. One uses Proposition 4.2 and the next result for this situation.

Proposition 4.3: Let $c(i,j)$ be symmetric in i and j. Then c is submodular if and only if (4.3) holds for $j \ge i$.

Proof: We must show that it also holds for $i > j$. By symmetry, (4.3) becomes for $i > j$:

$$c(j+1\ ,\ i+1)\ -\ c(j,i+1) \le c(j+1,i)\ -\ c(j,i) \tag{4.3)'}$$

If $i \ge j+1$, the result is true by hypothesis as the values of c in (4.3)' are all calculated above the main diagonal. But this is always the case as $i > j \Rightarrow i \ge j+1$.

Remark 4.2: If $c(i,j)$ is given only for $j \ge i$, we complete the matrix by symmetry. In (4.3)', for $i = j$ we must have: $c(i-1\ ,\ j-1)\ +\ c(i,j) \le c(i,j-1)\ +\ c(i-1,j)$ and $c(i,j-1)$ is below the main diagonal. It is therefore equal to $c(j-1,i) = c(i-1,j)$ as $i = j$. Thus, we must have:

$$c(i-1\ ,\ i-1)\ +\ c(i,i) \le 2c(i-1,i) \text{ for } i = 1,\ldots,n. \tag{4.4}$$

Example 4.2: Consider the cost function $c(i,j) = -i \cdot j\,|i-j|$, $i,j = 1,\ldots,n$. It is easy to check that$c(i,j)$ is a submodular function by checking (4.3) and (4.4). The case $j > i+1$ deals with the situation when all four cells involved in verifying (4.3) are above the main diagonal. In this case, we can verify (4.3) by looking at the extended function $c(x,y) = -xy(y-x)$, $y > x > 0$. The first derivative in x gives : $\frac{\partial c(x,y)}{\partial x} = -y^2 + 2xy$ and then $\frac{\partial^2 c(x,y)}{\partial y \partial x} = -2y + 2x = -2(y-x) < 0$. Thus submodularity holds. In order to check (4.4), we notice that the left hand side in (4.4) is 0 because the cost is 0 on the diagonal, due to the $|i-j|$ component. On the right hand side we have a positive function, thus (4.4) holds. As a particular case of Example 4.2, we have that the NWCR can be applied to give a basic feasible solution which is also optimal for Lahiri's problem. (Result 5.6.1, 5.6.2 of Arthanari and Dodge (1981)). More precisely, we have a simpler version of Example 4.2:

Example 4.3: The function $c : N^2 \to R_+$ defined by $c(i,j) = |i-j|$, $1 \leq i,j \leq N$ is submodular.

The optimality of Lahiri's solution viewed as a transportation problem is known. The following application to controlled sampling doesn't seem to have been presented in the literature. We consider the example given in Cochran (1977) p. 126, but one can equally apply it to the problem of Goodman and Kish (1950). This type of controlled sampling for two strata can be formulated as a transportation problem, the marginals being inclusion probabilities. We now show that the cost matrix is submodular and we actually apply the NWCR algorithm to obtain an optimal solution.

Example 4.4: **(see Example 3.7):** The following table represents the cost matrix associated with the problem. The revised order is used. Across, the numbers 1,...,5 represent the 5 hospitals of the small hospitals stratum and down are the large hospitals. The sampling plan gives the probabilities of selection that are used as marginals:

	1	2	3	4	5	$\pi(\ell)$
1	0	0	0	1	1	1/4
2	0	0	0	1	1	1/4
3	1	1	1	0	0	1/4
4	1	1	1	0	0	1/4
$\pi(s)$	1/5	1/5	1/5	1/5	1/5	

As before, $\pi(s)$ refers to the small hospitals and $\pi(\ell)$ to the large hospitals. The preferred samples were given a cost 0 and the non preferred a cost of 1. We recall that each sample consists of two units, one from each stratum. It is clear from the previous results that the cost function is submodular. Because of the symmetry, we only need to check (4.3) for $j \geq i$ (see Proposition 4.3) which is clearly satisfied here. Thus, on applying NWCR we will obtain an optimum solution. First, the cell with a minimum value for $i+j$ is selected, i.e. the cell (1,1). According to the algorithm, we must obtain q_{11} = min(1/4,1/5) = 1/5 = 0.2 and this represents the probability of selecting the sample (1,3'), where 1 is the first unit of the stratum of large hospitals and 3' belongs to the small hospitals.

We now eliminate column 1, as it gives the minimum probability and reset $\pi_1(\ell) = 1/4 - 1/5 = 1/20$. The new matrix is:

	2	3	4	5	$\pi(\ell)$
1	0	0	1	1	1/20
2	0	0	1	1	1/4
3	1	1	0	0	1/4
4	1	1	0	0	1/4
$\pi(s)$	1/5	1/5	1/5	1/5	

and we select now the cell with $\min(i+j)$ of the remaining cells. This, of course, is the cell (1,2) and $q_{12} = \min(1/20, 1/5) = 1/20 = 0.05$. We delete row 1 and reset $\pi_2(s) = 1/5 - 1/20 = 3/20 = 0.15$. The new table is:

	2	3	4	5	$\pi(\ell)$
2	0	0	1	1	1/4
3	1	1	0	0	1/4
4	1	1	0	0	1/4
$\pi(s)$	0.15	1/5	1/5	1/5	

Next, we select cell (2,2) and set $q_{22} = \min(1/4\ ,\ 0.15) = 0.15$. Column 2 is now deleted and $\pi_\ell(2)$ is reset at $1/4 - 3/20 = 0.1$. Continuing in this manner, we obtain the rest of the optimal solution: $q_{23} = 0.1$, $q_{33} = 0.1$, $q_{34} = 0.15$, $q_{44} = 0.05$, $q_{45} = 0.2$. The other probabilities are 0. Identifying the units in the original set-up and comparing with Cochran's solution, we see that the NWCR gives the same solution as his. Therefore, Cochran's solution is optimal. It can be shown in general that, for a given ordering of units, Lahiri's algorithm, which is the one used by Cochran (see p. 237 of Arthanari and Dodge) and the NWCR are one and the same. Both algorithms consists of two partitions of the interval [0,1] into subintervals of lengths equal to the marginals associated with rows, respectively columns. For the example above each $r \in [0,1]$ will fall into an interval $I_i(\ell)$ of length $\pi_i(\ell)$ for some $1 \leq i \leq 4$ and $I_j(s)$ of length $\pi_j(s)$ for some $1 \leq j \leq 5$. This means that the sample (i,j) receives a probability of selection equal to the intersection of the two intervals. If the intersection is empty, the sample has probability 0 of being selected and if $I_i(s)$ is included in $I_j(\ell)$, then $q_{ij} = \pi_i(s)$.

Clearly, the procedure as such can be generalized to n-dimensional matrices for n > 2. In order for the procedure to lead to an optimal solution, one has to find adequate orderings of the population so that the cost matrix is submodular.

From the previous example and discussion, it is clear that, if the problem can be modelled as a transportation problem and the cost matrix presents itself as blocks of 1's and 0's as in Example 4.4, we can treat the entire block as a unit and check submodularity among blocks. For example, the initial matrix in Example 4.4 could be represented by 4 entries:

	Type I	Type II	Stratum 1
Type II	0	1	
Type I	1	0	
Stratum 2			

Remark 4.3: The applicability of the NWCR in obtaining an optimal solution depends on a convenient labeling of units as seen in both Lahiri's and Goodman and Kish's examples. Once this is done and if the cost matrix becomes submodular, the selection of cells is made by starting first with the Northwest Corner, and then moving to the right along rows or down on columns to select a cell from the remaining cells.

A natural question arises: does Example 4.4 generalize to more strata or more types of units?

We restrict the discussion to the case of two strata for now. To generalize the model for 3 types of units when two units are chosen, one from each stratum and units of different types are preferred, we note first that, for any grouping of identical units and any labelling of types in one stratum, there are 3! possibilities for the other, of which some clearly do not lead to a submodular matrix, as the following example shows:

	T_1	T_2	T_3	Stratum 1
T_1	1	0	0	
T_2	0	1	0	
T_3	0	0	1	
Stratum 2				

Furthermore, for the T_1 , T_2 , T_3 ordering in stratum 1 if we start with T_2 and T_1 in stratum 2 we must have

$$\begin{array}{c|cc} & T_1 & T_2 \\ \hline T_2 & 0 & 1 \\ T_1 & 1 & 0 \end{array}$$

as a submatrix, but then placing T_3 last in both strata does not render the larger matrix submodular:

$$\begin{array}{c|ccc} & T_1 & T_2 & T_3 \\ \hline T_2 & 0 & 1 & 0 \\ T_1 & 1 & 0 & 0 \\ T_3 & 0 & 0 & 1 \end{array}$$

Therefore, for a specific ordering of types in stratum 1, stratum 2 must be more"thoroughly mixed", i.e. T_3 should be among the first two types. This reasoning, based on the fact that every submatrix of a submodular matrix must be submodular leads to the following conclusion:

Example 4.5: Consider the situation when two units must be chosen from two different strata. There are three types of units in each stratum and units of different types are preferred. We model this situation by assigning a cost of 0 to every sample that contains units of different types and 1 otherwise. It is then not possible to obtain a submodular cost by reordering the units.

An appropriate generalization can be obtained if three-way stratification is used, with one unit being selected per stratum and only different types of units forming a preferred sample.

To see this, we order the types of units as follows: along the x-axis, we set T_1, T_2, T_3, along the y-axis, T_3, T_1, T_2 and vertically T_2, T_3, T_1. Assume that we have three strata and we select three units, one from each stratum. We prefer samples that contain all types of units. We then give cost 0 to all such samples and 1 to the remaining samples. We have then an analogue of a two dimensional Monge matrix that has 0's in all diagonal blocks and 1 in all the other blocks. Therefore, it is possible, in some particular situations, to give ad-hoc solutions that allow the problem to be solved as a transportation problem, with a submodular cost function. As was the case with Cochran's hospital example, we can use a simple extension of Lahiri's selection scheme to obtain an optimum solution.

5. Sampling on one occasion. The structure of the optimum solution.

In this section, we restrict our attention to the simplest case of one survey in which samples of size $n = 2$ are selected, without replacement, from a population of size N. The results could be generalized straightforwardly to larger sample sized, but the statements are cumbersome. For instance, if the probability of a preferred sample of size n is increased, n, rather than two equations are affected and at least that many probabilities must be modified. As mentioned in Remark 3.5, the situation for one survey is more complex than that modelled by the transportation problem, since additional constraints appear as a result of identifying certain samples. Of course, all these problems could be solved using up-to-date versions of the simplex algorithm. However, this doesn't seem to be a very efficient way, even for small populations. The problem is that the number of possible samples and thus the number of variables is quite large. For example, for $N = 10$, there are 45 possible samples drawn without replacement. If only a few are preferred, many will be assigned a zero probability of selection in an optimal solution. It would be therefore useful to design a method that increases the number of zero variables among nonpreferred samples or decreases the cost function.

The problem that we study here is then similar to the problem studied by Rao and Nigam (1990). Consider a population P and let S be the set of all distinct samples considered for one survey. We study the solutions of (5.1), where S_1 is the set of all nonpreferred samples of S.

$$\text{(5.1)} \qquad \begin{aligned} &\sum_{\substack{s\in S \\ u\in s}} q(s) = \pi_u \,, \; all\ u \in P \\ &q(s) \geq 0 \qquad all\ s \in S \\ &\min_q \sum_{s\in S_1} q(s) \end{aligned}$$

Recall that samples are identified if they contain the same units. For example $s = (u_1 , u_2)$ is the same as (u_2 , u_1) in S. Therefore, for $n = 2$, $q(s)$ will appear in two equations, corresponding to π_{u_1} and π_{u_2}. As a consequence, we show that it is not always possible to assign maximum "mass" to a single preferred sample, in sharp contrast to the transportation problem.

<u>Example 5.1</u>: Let $P = \{1,2,3,4\}$, $n = 2$. We list the samples:
$s_1 = \{1,2\}$, $s_2 = \{1,3\}$, $s_3 = \{1,4\}$, $s_4 = \{2,3\}$, $s_5 = \{2,4\}$, $s_6 = \{3,4\}$.

Let $q_i = q(s_i)$, $i = 1,\ldots,6$. From (5.1), we have

$$\begin{aligned} q_1 + q_2 + q_3 &= \pi_1 \\ q_1 + q_4 + q_5 &= \pi_2 \\ q_2 + q_4 + q_6 &= \pi_3 \\ q_3 + q_5 + q_6 &= \pi_4 \end{aligned}$$

The preferred sample is s_1 and $\pi_1 = \min(\pi_1 , \pi_2)$. The question is, can we find a solution with $q_1 = \min\{\pi_1 , \pi_2\} = \pi_1$? The answer is not always affirmative. If we set $q_1 = \min(\pi_1 , \pi_2\} = \pi_1$ then $q_2 = q_3 = 0$ and the system becomes:

$$q_4 + q_5 = \pi_2 - \pi_1$$
$$q_4 + q_6 = \pi_3$$
$$q_5 + q_6 = \pi_4 \; .$$

As the determinant is different form zero, there is a unique solution given by: $q_4 = \dfrac{\pi_3 + \pi_2 - \pi_1 - \pi_4}{2}$, $q_5 = \dfrac{\pi_2 + \pi_4 - \pi_1 - \pi_3}{2}$, $q_6 = \dfrac{\pi_1 + \pi_3 + \pi_4 - \pi_2}{2}$. We also need the positivity constraints $q_i \geq 0$, $i = 4,5,6$. This leads to the following conditions on the probabilities of inclusions:

$$(5.2) \qquad |\pi_3 - \pi_4| \leq |\pi_2 - \pi_1| \leq \pi_3 + \pi_4 .$$

Unless these conditions are satisfied, we cannot set $q_1 = \min\{\pi_1 , \pi_2\}$. Notice that for simple random sampling without replacement condition (5.2) is satisfied. The last inequality is also satisfied because $\pi_1 + \pi_2 + \pi_3 + \pi_4 = 2$.

Next, we discuss the structure of the optimum solution. The following lemma is similar to Lemma 4.1. We do not assume that the cost function is submodular, although it only has values 0 and 1. We do not require any particular ordering of units either. What is required is the list of preferred samples S_0 and, consequently the list of nonpreferred samples S_1 , $S = S_1 \cup S_0$.

<u>Lemma 5.1</u>: For an optimum solution q to (5.1) and $s = (a,b) \in S_0$, we cannot have $q(a,x) > 0$ and $q(b,y) > 0$, where $(x,y) \in S$ and $(a,x) \in S_1$, $(b,y) \in S_1$

<u>Proof</u>: This is very similar to the proof of Result 5.6.1 of Arthanari and Dodge (1981) and will only be sketched here. Assume the conclusion does not hold. Let us then take $\varepsilon = \min(q(a,x) , q(b,y))$, $q^*(a,b) = q(a,b) + \varepsilon$, $q^*(a,x) = q(a,x) - \varepsilon$, $q^*(b,y) = q(b,y) - \varepsilon$, $q^*(x,y) = q(x,y) + \varepsilon$ for all the other $s \in S$. Then the system in (5.1) is verified by q^*, if q verifies it. The cost function corresponding to q^* is $c^* = c - \varepsilon < c$. If $(x,y) \in S_0$, then $c^* = c - 2\varepsilon < c$.

It has to be noted that the restriction $x \neq y$ is required when dealing with sampling without replacement. In this case samples of the type (x,x) are excluded from S.

<u>Example 5.2</u>: Let $P = \{1,2,3,4\}$ and consider all samples of size 2 listed in Example 5.1. From (5.1), we have with q a solution and $S_0 = (1,2)$:

$$\begin{aligned} q(1,2) + q(1,3) + q(1,4) &= \pi_1 \\ q(1,2) + q(2,3) + q(2,4) &= \pi_2 \\ q(1,3) + q(2,3) + q(4,3) &= \pi_3 \\ q(1,4) + q(2,4) + q(3,4) &= \pi_4 \quad . \end{aligned}$$

Assume that the feasible solution is such that $q(1,3)$, $q(1,4)$, $q(2,3)$, $q(2,4)$ are all larger than zero. We can start by taking $\varepsilon = \min\{q(1,3) \, , \, q(2,4)\} = q(1,3)$, say and change $q(1,2)$ to $q(1,2) + \varepsilon$, set $q^*(1,3) = 0$, $q^*(2,4) = q(2,4) - \varepsilon = q(2,4) - q(1,3)$, $q^*(3,4) = q(3,4) + \varepsilon$. The rule of indices is:

$$\begin{tikzcd} (1,2) \arrow[r] \arrow[d] & (3,1) \arrow[d] \\ (2,4) \arrow[r] & (3,4) \end{tikzcd}$$

with the upper left corner and the lower right corner receiving an increase in the cost functions and the others a decrease. Note $(1,2) \in S_0$, $(1,3)$, $(2,4) \in S_1$ and $(3,4) \in S$. Now the first two equations become:

$$\begin{aligned} q(1,2) + q(1,4) &= \pi_1 \\ q(1,2) + q(2,3) + q(2,4) &= \pi_2 \quad . \end{aligned}$$

We now pair $q(1,4)$ and $q(2,3)$ which will be decreased, and increase $q(1,2)$ as well as $q(3,4)$ in the third equation. Assume $q^*(2,3) = 0$. Now we have the system:

$$\begin{aligned} q(1,2) + q(1,4) &= \pi_1 \\ q(1,2) + q(2,4) &= \pi_2 \quad . \\ q(3,4) &= \pi_3 \\ q(1,4) + q(2,4) + q(3,4) &= \pi_4 \quad . \end{aligned}$$

This is an optimal solution. Any increase in $q(1,2)$ will require a decrease in $q(1,4)$ and $q(2,4)$. But then $q(3,4)$ should be decreased in the forth equation, which is impossible as its value is determined and equal to π_3.

The next result deals with an improvement of a solution in a case complementary to that of Lemma 5.1. We will then have a general characterization of an optimal solution by putting together Lemma 5.1 and 5.2.

Lemma 5.2.

Assume $N \geq 5$, $n = 2$, $(a,b) \in S_0$. For an optimal solution to (5.1), one cannot have $q(a,x) > 0$, $q(b,x) > 0$, $q(y,z) > 0$ with (a,x) , $(b,x) \in S_1$, $(y,z) \in S_1$, and (x,y) , $(x,z) \in S$.

Proof: Assume that the conclusion is not correct for a feasible solution q. Let $\varepsilon = \min\{q(a,x)\ ,\ q(b,x)\ ,\ q(y,z)\}$ and modify q as follows:

$$\begin{aligned} q^*(a,b) &= q(a,b) + \varepsilon \\ q^*(a,x) &= q(a,x) - \varepsilon \\ q^*(b,x) &= q(b,x) - \varepsilon \\ q^*(x,y) &= q(x,y) + \varepsilon \\ q^*(x,z) &= q(x,z) + \varepsilon \\ q^*(y,z) &= q(y,z) - \varepsilon \\ q^*(s) &= q(s) \quad \textit{for any other } s \in S \quad . \end{aligned}$$

It is clear that the equations corresponding to π_a and π_b in (5.1) are satisfied. Now (a,x) and (b,x) belong to the same equation and we choose two other existing samples containing, the unit x, say (x,y) and (x,z) and increase their probabilities to balance the equation corresponding to π_x. But now the left hand side of the equations corresponding to π_z and π_y have larger values for $q^*(x,z)$ and $q^*(x,y)$. These are offset by lowering $q(y,z)$ to $q^*(y,z)$. Thus, q^* is still a solution of (5.1) and again $c^* = c - \varepsilon$ as (y,z) , (a,x) and $(b,x) \in S_1$.

Example 5.3: Let $P = \{1,2,3,4,5\}$ and consider all possible samples of size $n = 2$, taken without replacement.

Let $S_0 = \{(1,2)\}$ and assume that Lemma 5.1 has been used repeatedly.

Then the system (5.1) reduces to:

$$\begin{aligned} q(1,2) + q(1,3) &= \pi_1 \\ q(1,2) + q(2,3) &= \pi_2 \\ q(1,3) + q(2,3) + q(3,4) + q(3,5) &= \pi_3 \\ q(3,4) + q(5,4) &= \pi_4 \\ q(3,,5) + q(5,4) &= \pi_5 \quad . \end{aligned}$$

Take $\varepsilon = \min(q(4,5)\ ,\ q(1,3)\ ,\ q(2,3))$ and decrease $q(4,5)$, $q(1,3)$, $q(2,3)$ by ε, increasing $q(1,2)$, $q(3,4)$ and $q(3,5)$ by ε. We thus create a 0 probability among the nonpreferred samples and reduce the cost function by ε. If $q^*(5,4) = 0$, we stop. But then $q^*(3,4) = \pi_4$, $q^*(3,5) = \pi_5$ and the system reduces to:

$$q(1,2) + q(1,3) = \pi_1$$
$$q(1,2) + q(2,3) = \pi_2$$
$$q(1,3) + q(2,,3) = \pi_3 - \pi_4 - \pi_5$$

which uniquely determines $q(1,2)$.

The solution thus obtained is optimal.

The following result puts together lemmas 5.1 and 5.2.

Theorem 5.1: Consider an optimal solution to (5.1) when the sample size Is $n = 2$. Then $\min\{q(a,x)\ ,\ q(b,y)\ ,\ q(s,t)\} = 0$, for all samples $(a,b) \in S_0$, (a,x) , (b,y) $(s,t) \in S_1$ and (x,s) , $(y,t) \in S$. If $(x,y) \in S$, then $\min\{q(a,x)\ ,\ q(b,y)\} = 0$.

Remark 5.1: The formulation and the proof of the above result point to the use of graph theory in solving problem (5.1) in the manner in which Arthanari and Dodge (1981) use it for transportation theory. It is also possible to give an algorithm like matrix interpretation of the mechanism of producing zeroes in the proofs of lemmas 5.1 - 5.2. In fact, the matrix approach can be easily generalized to sample sizes $n \geq 3$. This constitutes, however, the object of another study.

6. Concluding remarks

Although the general formulation of the problem in section 3 is quite broad, most of the discussion in this report is confined to the case of integrating 2 surveys or to controlled sampling when the sample size is 2. For such problems, algorithms other than the simplex algorithm are available. We investigated cases that could be set-up as transportation problems with submodular cost functions. In such instances, the Northwest Corner Rule algorithm can be used to give an optimum solution and Lahiri's method provides a sampling scheme that leads to the same solution.

D.A. Robertson pointed out that most of the examples that appear in the report could be treated as network flow problems, as the number of 1's on each column in the matrix of coefficients is 2. It would be interesting to investigate the efficiency of such algorithms and look for sampling schemes that would provide the same optimal solutions.

When the sample size is larger that 2 or when integrating more than 2 surveys, it would be interesting to investigate algorithms other than the simplex to obtain optimal solutions and find corresponding sampling schemes.

ACKNOWLEDGEMENT

This study was carried out while the author was on a Statistics Canada Sabbatical Programme. She would like to thank Professor J.N.K. Rao for having suggested the idea, Dr. D. Binder for the encouragement to participate in the programme and Dr. D.A. Robertson for many interesting comments.

REFERENCES

Arthanari, T.S. and Dodge, Y. (1981): Mathematical Programming in Statistics, Wiley, New-York.

Causey, B.D., Cox, L.H. and Ernst, L. (1985): Applications of transport theory to statistical problems. *JASA*, 80, 903-909.

Cochran, W.G. (1977): Sampling Techniques (ed. 3). Wiley, New-York.

Cox, L. and Ernst, L. (1982): Controlled rounding INFOR, 20, 423-432.

Fellegi, I.P. (1966): Changing the probabilities of selection when two units are selected with pps without replacement, Proceedings of the Social Statistics Section, American Statistical Assoc. 434-442.

Glover, F., Karney, Klingman, D. and Napier, A. (1974): A computation study on start procedures, basic change criteria and solution algorithm for transportation problems, Management Sciences 20, 793-813.

Goodman, R. and Kish, L. (1950): Controlled selection-a technique in probability sampling, *JASA*, 45, 350-372.

Hoffman, A.J. (1985): On greedy algorithms that succeed, Surveys in Combinations ed. I. Anderson, Camb. U. Press, 97-112.

Keyfitz, N. (1951): Sampling with probabilities proportionate to size: Adjustment for changes in probabilities, *JASA*, 46, 105-109.

Kish, L. and Scott, A. (1971): Retaining units after changing strata and probabilities, *JASA*, 66, 461-470.

Lahiri, D.H. (1954): Technical paper on some aspects of development of the sample design, *Sankhyā*, 14, 246- 316.

Mitra, S.K. and Pathak, P.K. (1984): Algorithm for optimal integration of three surveys, *Scan. J. Statis.*, 11, 257- 163.

Perkins, W.M. (1970): "1970 CPS Redesign: proposed method for deriving sample P.S.U. selection probabilities within 1970 N.S.R. strata, memorandum to Joseph Waksberg, U.S. Bureau of the Census.

Raj, D. (1957): On the method of overlapping maps in sample surveys, *Sankhyā*, 17, 89-98.

Rao, J.N.K. and Nigam, A.K. (1990): Optimal controlled sampling designs, *Biometrika*, 77, 807-814.

Rao, J.N.K. and Nigam, A.K. (1991): "Optimal" controlled sampling: a unified approach, preprint.

Ross, S.M. (1983): Introduction to Stochastic Dynamic Programming, Academic Press, San Diego.

Ca [illegible]