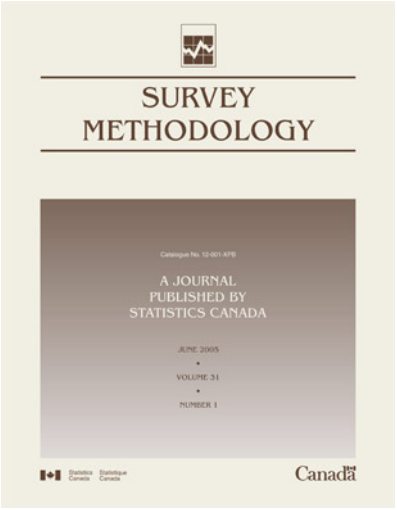
Catalogue no. 12-001-XIE

Survey Methodology

December 2002



Statistics Canada Statistique Canada

Canada

How to obtain more information

Specific inquiries about this product and related statistics or services should be directed to: Business Survey Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6 (telephone: 1 800 263-1136).

For information on the wide range of data available from Statistics Canada, you can contact us by calling one of our toll-free numbers. You can also contact us by e-mail or by visiting our website.

National inquiries line	1 800 263-1136
National telecommunications device for the hearing impaired	1 800 363-7629
Depository Services Program inquiries	1 800 700-1033
Fax line for Depository Services Program	1 800 889-9734
E-mail inquiries	infostats@statcan.ca
Website	www.statcan.ca

Information to access the product

This product, catalogue no. 12-001-XIE, is available for free. To obtain a single issue, visit our website at www.statcan.ca and select Our Products and Services.

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner and in the official language of their choice. To this end, the Agency has developed standards of service that its employees observe in serving its clients. To obtain a copy of these service standards, please contact Statistics Canada toll free at 1 800 263-1136. The service standards are also published on www.statcan.ca under About Statistics Canada > Providing services to Canadians.

Statistics Canada
Business Survey Methods Division

Survey Methodology

December 2002

Published by authority of the Minister responsible for Statistics Canada

January 2006

Catalogue no. 12-001-XIE
ISSN 1492-0921

Frequency: semi-annual

Ottawa

Cette publication est disponible en français sur demande (nº 12-001-XIF au catalogue).

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued cooperation and goodwill.

Vol. 28, No. 2, pp. 199-207
Statistics Canada, Catalogue No. 12-001

Multi-way Stratification by Linear Programming Made Practical

Wilson Lu and Randy R. Sitter [1]

Abstract

Sitter and Skinner (1994) present a method which applies linear programming to designing surveys with multi-way stratification, primarily in situations where the desired sample size is less than or only slightly larger than the total number of stratification cells. The idea in their approach is simple, easily understood and easy to apply. However, the main practical constraint of their approach is that it rapidly becomes expensive in terms of magnitude of computation as the number of cells in the multi-way stratification increases, to the extent that it cannot be used in most realistic situations. In this article, we extend this linear programming approach and develop methods to reduce the amount of computation so that very large problems become feasible.

Key Words: PPS sampling; Proportional allocation; Random grouping; Survey sampling.

1. Introduction

In many practical survey situations, there are multiple stratifying variables available and thus the designer has the option of defining strata as cells formed as cross-classified categories of these variables. For examples, see Engle, Marsden and Pollock (1971), Hess, Riedel and Fitzpatrick (1976), Vihma (1981) and Skinner, Holmes and Holt (1994). This multi-way stratification often leads to situations where the desired sample size is less than or only slightly larger than the total number of stratification cells (particularly common when choosing primary sampling units (psu's) in stratified multi-stage designs) and hence conventional methods of sample allocation to strata may not be applicable.

An illustration, based on a hypothetical example of Bryant, Hartley and Jessen (1960), is given in Table 1. Communities (psu's) are classified by two stratifying factors, type and region, with three and five categories respectively. The desired sample size of $n = 10$ is less than the total number of cells, 15. This example also illustrates a related problem. The entries in Table 1 are the expected counts under proportional stratification, *i.e.*, the strata sample sizes are proportional to the population strata sizes. Under the sample size restrictions, the expected cell sample counts will not generally be integers. In cases with very small expected counts, rounding to integers will not lead to good choices while causing a serious violation of the property of proportional allocation. Non-integer margin totals are also typical and can cause their own difficulties. Goodman and Kish (1950) was the first to address this problem under the name of controlled selection, where they propose a sampling selection procedure which can be classified as random sytematic sampling (see Hess, Riedel and Fitzpatrick 1976; Waterton 1983). Bryant *et al.* (1960) presented a very simple method to randomly assign sample sizes for each cell in two-way stratification and gave two estimators based on that sampling scheme. However, since the expected cell sample sizes didn't include information of proportion of each cell (*i.e.*, the method is not a proper controlled selection technique, as only the probabilities of the marginal distributions are respected), these estimators may not have satisfactory MSE properties (see Sitter and Skinner 1994). Jessen (1970) points out that a further limitation of the method of Bryant *et al.* (1960) is that it is not possible to constrain specified cell sizes to be zero, which may be desired in some situations (see related methods under the label "lattice sampling", *e.g.* Jessen 1973, 1975). He proposes two methods for both two-way and three-way stratification but both methods are fairly complicated to implement and, as noted by Causey, Cox and Ernst (1985), may not lead to a solution. Inspired by the idea of Rao and Nigam (1990, 1992) in the context of avoiding undesirable samples (see also Lahiri and Mukerjee 2000), Sitter and Skinner (1994) proposed a linear programming approach which attempts to take advantage of the power of modern computing. This linear programming technique is simple in conception, is flexible to different situations, always has a solution and has better properties of the MSE. Its main practical constraint is that it becomes computationally intensive as the number of cells in the multi-way stratification increases, quickly to the point of infeasibility. In this paper we will present a simple method which will allow the linear programming technique to handle much larger problems. In section 2 we describe the linear programming method of Sitter and Skinner (1994) to introduce notation and briefly discuss its numerical limitations. In section 3.1, we first discuss some simple strategies to reduce the computational intensity of the method as motivation for the eventual proposal. In sections 3.2 and 3.3 we discuss the proposed method assuming integer margins

1. Wilson Lu, Doctoral Student, Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, BC, Canada V5A 1S6; Randy R. Sitter, Professor, Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, BC, Canada V5A 1S6.

and give some examples with from 80 to 300 stratification cells to illustrate the ability of the new methodology to handle large problems. In section 3.4, we describe the simple extention of the method to non-integer margins and illustrate by applying the method to a real example from the occupational health literature (Vihma 1981).

Table 1
Example from Bryant *et al.* (1960). Expected Sample Cell Counts Under Proportional nStratification (n = 10)

		Type of Community		
Region	Urban	Rural	Metropolitan	Total
1	1.0	0.5	0.5	2.0
2	0.2	0.3	0.5	1.0
3	0.2	0.6	1.2	2.0
4	0.6	1.8	0.6	3.0
5	1.0	0.8	0.2	2.0
Total	3.0	4.0	3.0	10.0

2. The Linear Programming Technique

2.1 The Basic Ideas

We introduce the linear programming method of Sitter and Skinner (1994) by considering the simplest kind of twoway stratification. Suppose that N units of a finite population are arranged in a two-way classification in R rows formed by categories of one variable and C columns by categories of another. Let N_{ij} denote the number of population units in the i^{th} row and the j^{th} column (*i.e.*, in the ij^{th} cell) of the two-way table and $P_{ij} = N_{ij}/N$ denote the proportion of the total population in the ij^{th} cell. Let $\bar{Y}$ denote the mean value of a survey characteristic y for the population and $\bar{Y}_{ij}$ denote the mean value of y for the ij^{th} cell.

The sample is selected as follows:

i) Sample sizes n_{ij} are randomly determined for each cell according to a pre-specified procedure. Letting s denote the $R \times C$ array $(n_{ij},\ i = 1,\ ...,\ R,\ j = 1,\ ...,\ C)$, this procedure assigns a probability $p(s)$ to each s in the set S of possible such arrays and selects a single array, s, from S. We denote the dependence of n_{ij} on s by writing $n_{ij}(s)$.

ii) A simple random sample of $n_{ij}(s)$ units is then selected from the ij^{th} cell and the values of y obtained.

Restrict attention to designs of fixed sample size $n > 0$, that is, restrict to arrays $s \in S_n$ such that $\sum_{i=1}^{R}\sum_{j=1}^{C} n_{ij}(s) = n$. We would also like to restrict attention to proportionate stratification so that

$$\sum_{s \in S_n} n_{ij}(s)\ p(s) = nP_{ij} \text{ for } i = 1,\ \ldots,\ R,\ j = 1,\ \ldots,\ C, \quad (1)$$

which implies that the simple unweighted sample mean $\bar{y}(s)$ is an unbiased estimator of $\bar{Y}$. We will refer to (1) as the expected proportional allocation (EPA) constraint.

The linear programming technique of Sitter and Skinner (1994) chooses a sampling design $p(s)$ which minimizes the expected lack of 'desirability' of the samples by solving the linear programming problem:

$$\min \sum_{s \in S_n} w(s)\ p(s) \quad (2)$$

subject to the constraint (1), where $w(s)$ is a loss function for the sample s, to be specified, and the $p(s)$ are the unknowns. Sitter and Skinner (1994) were exploiting the key observation of Rao and Nigam (1990, 1992) in the context of avoiding undesirable samples, that the objective function in (2) was linear in the $p(s)$'s (see also Lahiri and Mukerjee 2000).

In the objective function (2), the loss function $w(s)$ plays an important role. With a well defined $w(s)$, we have flexibility to explore the existence of an optimal solution to (2) within an economically sized S_n and, more importantly, to improve efficiency of estimation. Sitter and Skinner (1994) suggest choosing

$$w(s) = \sum_{i=1}^{R} \left(n_{i.}(s) - nP_{i.}\right)^2 + \sum_{j=1}^{C} \left(n_{.j}(s) - nP_{.j}\right)^2, \quad (3)$$

where $n_{i.}(s) = \sum_j n_{ij}(s), n_{.j}(s) = \sum_i n_{ij}(s), P_{i.} = \sum_j P_{ij}$ and $P_{.j} = \sum_i P_{ij}$. Obviously, the objective function (2) is actually $E(w(s))$ for any given design $p(s)$ and can be explained as the mean squared error of estimator $\bar{y}$ under an analysis of variance model (see Sitter and Skinner 1994). Then by solving the above linear programming problem, one can obtain minimized MSE in the sense of ANOVA while maintaining the EPA property of the $n_{ij}(s)$. One should note that if a design with objective function equal to zero is obtained, then all margin constraints are met. This would typically only be the case with integer margins.

Sitter and Skinner (1994) suggest that one simple way to reduce the size of S_n is to restrict the actual values that n_{ij} can take to be either $\lfloor nP_{ij} \rfloor$ or $\lfloor nP_{ij} \rfloor + 1$, where $\lfloor nP_{ij} \rfloor$ is the greatest integer less than or equal to nP_{ij}. By denoting $\tilde{n}_{ij} = n_{ij} - \lfloor nP_{ij} \rfloor$ and $r_{ij} = nP_{ij} - \lfloor nP_{ij} \rfloor$, one can then impose.

$$E(\tilde{n}_{ij}) = r_{ij}, \quad (4)$$

where $\tilde{n}_{ij} = 0$ or 1 and $0 \le r_{ij} < 1$. Then the linear programming method can be applied to the $\tilde{n}_{ij}$ and finally $\lfloor nP_{ij} \rfloor + \tilde{n}_{ij}$ can be used as the actual cell sample sizes. Therefore, without loss of generality, we will assume that

$$n_{ij} = 0,1 \text{ and } 0 \le r_{ij} = nP_{ij} < 1. \quad (5)$$

2.2 Higher-way Stratification

The Sitter and Skinner (1994) approach extends straightforwardly to more stratifying factors by letting s denote the corresponding r-way array. The loss function would then include more terms, for example for three-way stratification equation (3) could be replaced by

$$w(s)=\gamma_1\sum_{i=1}^{R_1}\left(n_{i..}(s)-nP_{i..}\right)^2+\gamma_2\sum_{j=1}^{R_2}\left(n_{.j.}(s)-nP_{.j.}\right)^2+\gamma_3\sum_{k=1}^{R_3}\left(n_{..k}(s)-nP_{..k}\right)^2$$

in obvious notation, where γ_1, γ_2 and γ_3 might represent the relative importance of balancing on the three factors based on prior information (see Sitter and Skinner 1994).

2.3 Multi-stage Sampling

An important application of multi-way stratification is to the selection of primary sampling units (psu's) in multistage sampling, where it is more common to have several stratifying factors available.

In section 2.1, the inclusion probabilities of each unit are $E(n_{ij}(s)/N_{ij})=n/N$. If psu's are selected with equal probability then the approach extends directly with the psu's the units and with the observed values of y replaced by unbiased estimators of the psu totals. However, if the psu's are to be selected with unequal probabilities, say nz_{ijk} for psu k in stratification cell ij (z_{ijk} will typically equal $M_{ijk}/\sum_{ijk}M_{ijk}$, with M_{ijk} being some measure of size of psu k in cell ij), then the procedure can be easily modified by setting P_{ij} equal to $z_{ij.}/z...$, where $z_{ij.}=\sum_k z_{ijk}$ and $z...=\sum_{ijk}z_{ijk}$. Then, if $n_{ij}(s)>0$, a sample of $n_{ij}(s)$ psu's in cell ij is selected by some probability proportional to z_{ijk} method.

2.4 An Example

The linear programming approach can be illustrated using the hypothetical example of Bryant *et al.* (1960) given in Table 1. First, this problem is simplified as shown in Table 2 to meet the assumption in (5). Then, a standard linear programming package is used to solve this reduced problem (2). Because integer margins of expected sample cell counts can be exactly matched by marginal totals of sample sizes $n_{i.}$ and $n_{.j}$, which means that the loss function $w(s)$ can acheive a minimum value of zero, the objective function in (2) for this example is also minimized at zero. The optimal solution of this problem is given in Table 3. It should be noted that this solution has been converted back to match the original example shown in Table 1.

Table 2
Modified Example from Bryant *et al.* (1960)

	Type of Community			
Region	Urban	Rural	Metropolitan	Total
1	0.0	0.5	0.5	1.0
2	0.2	0.3	0.5	1.0
3	0.2	0.6	0.2	1.0
4	0.6	0.8	0.6	2.0
5	0.0	0.8	0.2	1.0
Total	1.0	3.0	2.0	6.0

Table 3
Linear Programming Solution to Example from Bryant *et al*. (1960)

s	$p(s)$	s	$p(s)$	s	$p(s)$
1 1 0		1 1 0		1 1 0	
1 0 0		0 0 1		0 0 1	
0 1 1	0.2	0 1 1	0.1	0 0 2	0.2
0 2 1		1 1 1		1 2 0	
1 0 1		1 1 0		1 1 0	
1 0 1		1 0 1		1 0 1	
0 1 0		0 1 0		0 0 1	
0 1 0	0.2	0 1 1	0.1	0 1 1	0.2
0 2 1		1 1 1		1 2 0	
1 1 0		1 1 0		1 1 0	

The linear programming method is simple and easy to use. Its main drawback is computational. The number of parameters in the resulting linear programming problem is the number of samples of size n from the $RC>n$ cells, $\binom{RC}{n}$, which becomes infeasibly large quite quickly. In the next section we will explore ways of improving the computational efficiency of the linear programming approach while maintaining all of its good properties.

3. The Linear Programming Approach Made Practical

The basic idea of the linear programming approach is to obtain an optimal sampling design in terms of the (minimum) expected lack of "desirability" of the sample by directly solving a linear programming problem with $p(s), s\in S_n$, as the unknowns while maintaining the EPA property. The only obstacle to this approach is that the number of elements in S_n is often very large and even with modern computing power it becomes difficult to carry out linear programming if the number of unknowns is large.

To reduce the magnitude of the computational task for this linear programming problem determined by the cardinality of S_n, we want to obtain a subset of S_n, say S_{n0}, which is nearly as representative as S_n but much smaller, and thus solve the following linear programming problem with a much smaller set of $p(s), s\in S_{n0}$, as the unknowns:

$$\min \sum_{s\in S_{n0}} w(s)p(s). \qquad (6)$$

Hopefully, in this way we can easily deal with larger practical problems without losing the good properties of the linear programming approach.

3.1 Some Motivating Strategies

The above strategy is easy to state, but it turns out not to be entirely obvious how to go about it. In fact, there are several different directions we can explore to determine such a subset $S_{n0}\subset S_n$. In this section, we will describe a

basic method related to loss functions which was alluded to in Sitter and Skinner (1994) and describe how it modestly increases the size of problems that can be handled. We will then discuss some obvious directions to take which did not improve things much. By describing these misguided attempts, we motivate the eventual proposal.

The major flexibility of the linear programming approach is derived from the choice of loss function $w(s)$. Thus, it is natural for us to consider the loss function first when we try to improve the computational efficiency of this approach. By observing the objective function of the linear programming problem (2), we suspect that the loss function $w(s)$ as coefficients of unknowns $p(s)$ will not be very large when the objective function has been minimized. In other words, all positive $p(s)$ in an optimal sampling design will only be assigned to samples having small lack of "desirability". Based on this observation, we hypothesize that the following subset might be a good replacement for S_n,

$$S_{n0} = \left\{ \begin{array}{l} s \in S_n : w(s) = \sum_{i=1}^{R} \left(n_{i.}(s) - nP_{i.}\right)^2 \\ + \sum_{j=1}^{C} \left(n_{.j}(s) - nP_{.j}\right)^2 \le w_0 \end{array} \right\}, \qquad (7)$$

where w_0 is a pre-determined positive constant. In the case of integer margins, one could even let $w_0 = 0$ and restrict to samples where the margins are matched. For example, the solution in Table 3 assigned positive probability to only 6 samples and for each of these the objective function was zero.

Lu (2000) develops nested linear programming strategies for solving this problem. For moderately sized problems such as 8×5 arrays (*i.e.*, 40 cells) this approach does well. However, for larger problems the size of resulting candidate sets becomes large very quickly, even in the integer margin case. Thus for large problems the technique faces the same problem as before-a huge candidate set that results in the difficulty of solving a linear programming problem with too many unknowns.

In reality, even a candidate sample set S_{n0} of the form in (7) is far larger than necessary for us to find an optimal solution. What we really need is a smaller but fairly representative subset, where by "small" we mean small enough to make it *possible* to solve the resulting linear programming problem and by "representative" we mean containing elements which promise that this linear programming problem is *feasible*.

Before going on to describe our eventual proposed solution to this problem, we would like to introduce some naive methods of obtaining such a "representative subset" that turned out not to work well. These are not that useful in practice, but they did inspire our thinking in proposing a more sophisticated approach.

1) Two Stage Optimization: First of all, we could try to break S_{n0} in (7) into many subsets which are small enough to be handled by linear programming respectively. Hopefully, optimal solutions from each of these smaller sets in the first stage optimization procedure can be combined to form the desired representative set of samples. Then we can just collect these optimal solutions together and apply linear programming once more. We applied this method to some simulated examples of size 6×6, 7×7, 8×8 and 9×9 as a method of preliminary investigation of its potential. Generally, in the first two cases the method worked very well and quickly, in the 8×8 case the method was time consuming and was not always able to obtain optimal solutions, and in the 9×9 case the method became infeasible.

2) Resampling from S_{n0}: We could also randomly select a proportion, say 10%, of the S_{n0} in (7) and hope this proportion is statistically representative of the complete set. Unfortunately, simulation results showed that the proportion obtained in this way is not "representative" enough, and the resulting linear programming problem often does not have any feasible solution. For example, the method of nested linear programming discussed previously was able to obtain matched integer margin solutions for simulated 8×5 arrays, however, these solutions were obtained much quicker by repeatedly sampling 10% of S_{n0} and applying the Sitter and Skinner (1994) method to this set until a feasible solution was obtained. However, when slightly larger cases were considered the method took an inordinate amount of time before finding a feasible solution, and quickly became impractical.

There are two problems with both these approaches. First, the size of S_{n0} becomes huge combinatorically and even complete enumeration becomes difficult. Having to first obtain S_{n0} and then cutting the problem into pieces will either quickly outstrip the practical limits on linear programming due to the size of the pieces or create a huge number of pieces. Second, both of these strategies are not in any way attempting to avoid samples which are particularly bad choices for meeting the EPA constraints. The question is, is there any way we can generate a fairly "representative" candidate sample subset without choosing such "useless" samples or, more generally, can we select candidate samples in which the frequency of an entry's appearance is more or less related to its desired expected sample counts?, and also can we do so without first having to enumerate a large S_{n0} ? The general idea revolves around the fact that if we could randomly select a candidate subset directly from S_n without complete enumeration using an unequal probability selection procedure which simultaneously ensures that the objective function is minimized for every sample while ensuring that the EPA property is satisfied we will have solved the problem without resorting to linear programming at all. We have been working on finding such a selection procedure, but have yet to succeed. What we have been able to do is to develop such a procedure

with approximate EPA (AEPA). We can then use it to randomly generate a candidate subset of samples, S_{n0}, and then apply a linear programming technique to this subset.

3.2 A Sampling Procedure with AEPA Property

In this section we first describe the approach as it applies to the case of integer margins. That is, the column totals, $n_{.j} = \sum_{i=1}^{R} r_{ij}$, and the row totals, $n_{i.} = \sum_{j=1}^{C} r_{ij}$, are integer valued. We go on to discuss how it can easily be adapted to the general case. In the linear programming approach, the goal is to minimize the expected lack of 'desirability' of the samples while maintaining the EPA property. We propose to accomplish this in two stages. First, we will develop an unequal probability selection procedure which selects samples which exactly match the integer margins and also have the AEPA property. We will then randomly generate a moderately sized set of such arrays and then apply a modified linear programming technique to this subset of all possible arrays. This will be repeated with larger and larger such sets. We will describe the sampling procedure and then we will discuss the modified linear programming technique.

Here is the basic idea for constructing such a sampling procedure: for a two-way table (assuming the expected cell sample sizes have been adjusted to lie between 0 and 1 as was done in going from Table 1 to 2), first we draw a sequence of population cells to produce $a_{11}, a_{12}, ..., a_{1C}$ in the first row using an unequal probability without replacement sampling procedure based on the expected counts of that row, where $a_{ij} = 1$ if the ij^{th} cell is selected and = 0 otherwise.Then we draw $a_{i1}, a_{i2}, ..., a_{iC}$ subsequently for $i > 1$ while keeping all $\sum_{k=1}^{i} a_{kj}$ less than or equal to the corresponding marginal column totals n_j. The details of this sampling procedure are as follows:

Step 1: Randomly permute the rows and let $i = 1$. Given the first row of inclusion probabilities $r_{11}, r_{12}, ..., r_{1C}$, draw a sample of $n_{1.}$ cells out of C in the first row stratum using an unequal probability without replacement sampling procedure; record the first row of samples in terms of indicator variables $a_{11}, a_{12}, \ldots, a_{1C}$ as defined previously; let $A_j = a_{1j}$ for $j = 1, \ldots, C$.

Step 2: Let $i = i + 1$.

Step 2.1: For $j = 1, \ldots, C$, do the following:

a) Let $R_j = \sum_{k=1}^{i} r_{kj}$,

b) If $R_j - A_j \le 0$ let $a_{ij} = 0$,

c) If $R_j - A_j \ge 1$ let $a_{ij} = 1$.

Step 2.2: Let $J = \{j : 0 < R_j - A_j < 1\}$ and $rtot = \sum_{j=1}^{C} r_{ij} - \#\{j : a_{ij} = 1\}$. If $rtot > 0$ then $r'_{ij} = r_{ij} \times rtot / \sum_{j \in J} r_{ij}$, for $j \in J$. If there exists a $j_0 \in J$ such that $r'_{ij0} > 1$, then let $a_{ij0} = 1$ and go to Step 2.1. Otherwise go to Step 3.

Step 3: Draw a sample of *rtot* cells from J using an unequal probability without replacement sampling procedure and r'_{ij} to get a_{ij} for $j \in J$.

Let $A_j = \sum_{k=1}^{i} a_{kj}$ for $j = 1, \ldots, C$.

Step 4: If $i = R$, then stop; otherwise go to Step 2.

One aspect of this sampling procedure that should be noticed is that in Step 2, the way of re-calculating the i^{th} row of inclusion probabilities is not unique. However, the general rules that should be followed for this re-calculation are:

a) $0 \le r'_{ij} \le 1$ and if $A_j = n_{.j}$, which means that there are enough units being selected from the j^{th} column, r'_{ij} should be set to 0; if $A_j = n_{.j} - (R - i + 1)$, which means that there will not be enough units to be selected for this column unless all of the remaining units are selected, r'_{ij} should be set to 1;

b) keep $\sum_{j=1}^{C} r'_{ij} = \sum_{j=1}^{C} r_{ij} = n_{i.}$.

The method extends easily to non-integer margins. We delay detailed discussion, however, to the sequel.

We can now use the above method to generate a candidate set, S_{n0}, and apply the linear programming technique to this set. To see why we choose to modify the linear programming technique, realize that for the integer margin case every $s \in S_{n0}$ already attains the minimum in (2) so that a direct application of linear programming amounts to determining whether there is a feasible solution or not. Thus, if we generate say an S_{n0} of size 500 then 1 000 *etc*, and the linear programming package continues to find no feasible solution we really do not know if we are getting closer to a solution or not. Instead we choose to turn the optimization around and solve a dual problem

$$\min_{p(s)} \sum_{i,j} \left| \sum_{s \in S_{n0}} n_{ij}(s) p(s) - r_{ij} \right|. \tag{8}$$

We know that $w(s) = 0$ for all $s \in S_{n0}$ and we are looking for a solution which yields a minimum of zero in (8). We have essentially switched the roles of the objective function and the EPA constraints in the original problem. The difficulty is that it is more difficult to use linear programming to handle (8). This can be done as follows. Set up constraints

$$\sum_{s \in S_{n0}} n_{ij}(s) p(s) - r_{ij} + d_{ij} - e_{ij} = 0 \quad \text{for } i = 1, \ldots, R \text{ and } j = 1, \ldots, C, \tag{9}$$

where

$$d_{ij} \ge 0, e_{ij} \ge 0, d_{ij} e_{ij} = 0. \tag{10}$$

Then note that

$$\left| \sum_{s \in S_{n0}} n_{ij}(s) p(s) - r_{ij} \right| = \begin{cases} d_{ij} \text{ if } \sum_{s \in S_{n0}} n_{ij}(s) p(s) - r_{ij} < 0 \\ e_{ij} \text{ if } \sum_{s \in S_{n0}} n_{ij}(s) p(s) - r_{ij} \ge 0 \end{cases} = d_{ij} + e_{ij}. \tag{11}$$

Thus, we can replace (8) by

$$\min_{p(s), d_{ij}, e_{ij}} \sum_{i,j} \left(d_{ij} + e_{ij} \right), \tag{12}$$

subject to

$$\sum_{s\in S_{n0}} n_{ij}(s)p(s) - r_{ij} + d_{ij} - e_{ij} = 0, d_{ij},$$

$$e_{ij}, p(s) \geq 0, d_{ij}e_{ij} = 0. \qquad (13)$$

3.3 Some Illustrating Examples with Integer Margins

In this section, two examples will be used to illustrate the sampling procedure. The first with a 10×8 array is described in detail to show the whole procedure. The second with a larger size (20×15) is given to demonstrate the size of problem that this method can handle (this is near the limit of the problem the proposed method can realistically handle). Any unequal probability without replacement sampling procedure can be used within the method. In Example 1 below, we chose to use the the random grouping method of Rao, Hartley and Cochran (1962), since it is simple and we really only need to approximately match the selection probabilities, which it does. However, the Rao-Hartley-Cochran method only works well up to problems of moderate size. In Examples 2 and 3 one should use a method which exactly matches the selection probabilities. There are many such available, but we chose to use one developed in Lu (2000).

Example 1. 10×8 array with integer margins: A two-way stratification problem with expected sample cell counts and sample size is given in Table 4.

Table 4
Expected Sample Cell Counts Under Proportionate Stratification (n = 40)

Row No.	Column No.								Marginal Row Total
	1	2	3	4	5	6	7	8	
1	0.41	0.55	0.58	0.80	0.23	0.61	0.70	0.12	4
2	0.52	0.15	0.07	0.90	0.28	0.10	0.37	0.61	3
3	0.72	0.15	0.65	0.73	0.39	0.34	0.85	0.17	4
4	0.70	0.55	0.46	0.10	0.41	0.05	0.24	0.49	3
5	0.07	0.63	0.45	0.81	0.52	0.02	0.70	0.80	4
6	0.61	0.33	0.79	0.21	0.02	0.61	0.67	0.76	4
7	0.88	0.48	0.73	0.69	0.,44	0.64	0.86	0.28	5
8	0.22	0.14	0.85	0.37	0.69	0.45	0.49	0.79	4
9	0.85	0.44	0.80	0.76	0.31	0.71	0.60	053	5
10	0.02	0.58	0.62	0.63	0.71	0.47	0.52	0.45	4
Marginal Col. Total	5	4	6	6	4	4	6	5	40

The basic steps of our sampling design are illustrated as follows:

Step 1. Obtain a representative candidate sample subset S_{n0} by using proposed sampling procedure with AEPA property to draw, say 500, samples (obtained within 3 minutes). The sample proportion of each cell is shown in Table 5, which can be compared to Table 4 to see how close these are to satisfying the EPA property.

Step 2. Solve the linear programming problem given by (12) and (13) to obtain

$$\min_{p(s), s\in S_{n0}} \sum_{i,j} \Big| \sum_{s} n_{ij}(s)p(s) - nP_{ij} \Big|. \qquad (14)$$

If the objective value of (14) is greater than zero, repeat Step 1 with a larger set S_{n0}. If the objective value of (14) is zero, stop, an optimal solution has been obtained.

Table 5
Sample Cell Counts Under Prop. Stratification (n = 40)

Row No.	Column No.								Marginal Row Total
	1	2	3	4	5	6	7	8	
1	0.408	0.554	0.582	0.776	0.250	0.594	0.734	0.102	4
2	0.554	0.150	0.062	0.916	0.280	0.122	0.366	0.550	3
3	0.690	0.144	0.638	0.720	0.402	0.360	0.838	0.208	4
4	0.692	0.542	0.452	0.120	0.416	0.044	0.260	0.474	3
5	0.060	0.602	0.446	0.814	0.568	0.016	0.708	0.786	4
6	0.558	0.348	0.780	0.216	0.012	0.634	0.682	0.770	4
7	0.866	0.480	0.734	0.676	0.470	0.664	0.842	0.268	5
8	0.254	0.158	0.848	0.400	0.654	0.412	0.490	0.784	4
9	0.870	0.418	0.830	0.772	0.292	0.692	0.624	0.502	5
10	0.026	0.564	0.636	0.658	0.714	0.416	0.500	0.486	4
Marginal Col. Total	5	4	6	6	4	4	6	5	40

In this example, a candidate subset S_{n0} with 500 samples was sufficient to get objective value of 0.

Example 2. 20×15 array with integer margins: In this example, a 20×15 array with integer margins is given in Table 6.

The actual computation steps are given as follows:

First Iteration:

Step 1. Draw 500 samples to form S_{n0}.
Step 2. The objective value of (14) is 0.1659.

Second Iteration:

Step 1. Draw 500 samples to add to S_{n0}.
Step 2. The objective value of (14) is 0. The final sampling design is attained.

This procedure took approximately 30–60 seconds using a Fortran program on a Sun Ultra 10 workstation.

3.4 Extension to Non-Integer Margins

The method extends easily to non-integer margins. Merely replace $n_{i.}$ throughout the algorithm by $n_{i.}^{*}$ which takes value $\lfloor r_{i.} \rfloor + 1$ with probability $\alpha = r_{i.} - \lfloor r_{i.} \rfloor$ and takes value $\lfloor r_{i.} \rfloor$ with probability $1-\alpha$. The only additional difficulty is that $E[w(s)]$ cannot attain zero. Thus, we do not have an obvious lower-bound reference point to ascertain whether we are close to the best solution or not. However, the above randomization strategy ensures that for every obtained AEPA sample we have

$$|n_{i.}(s) - r_{i.}| < 1 \text{ and } |n_{.j}(s) - r_{.j}| < 1 \quad \text{for } i = 1, \ldots, R, \ j = 1, \ldots, C. \tag{15}$$

This together with the EPA property, $E[n_{ij}(s)] = \sum_s n_{ij}(s)p(s) = r_{ij}$ implies that the lack of desirability function $w(s)$ defined in (3) has a constant expectation

$$E[w(s)] = \sum_i \left(r_{i.} - \lfloor r_{i.} \rfloor\right)\left(1 + \lfloor r_{i.} \rfloor - r_{i.}\right) + \sum_j \left(r_{.j} - \lfloor r_{.j} \rfloor\right)\left(1 + \lfloor r_{.j} \rfloor - r_{.j}\right). \tag{16}$$

The proof of this is given in Appendix 1. Thus, if (14) attains zero under the above strategy then the resulting solution will yield minimum $E[w(s)]$ as in (16).

Example 3. 27×3 real example with non-integer margins: We will illustrate the method using a real example from environmental health (Vihma 1981). This study was concerned with occupational health of workers in various industries in Finland. The population chosen for study consisted of 1,430 small industrial workplaces (5–49 employees) totalling 22,893 employees in Uusimaa, the southern most and most industrialized province of Finland. The primary sampling units were the workplaces and a sample of n = 100 such were desired. This was all that could be afforded given the cost of the eventual survey. The workplaces were stratified by two stratification variables: type of industry (27 categories) and number of employees (3 categories). The expected sample cell counts under proportionate stratification are given in Table 7. The actual sampling scheme used in this study was based on the method of Bryant *et al.* (1960) after some grouping strata as it was the only method available at the time of this study.

We applied our method to this problem. The minimum achievable $E[w(s)]$ using our proposed strategy is 5.0418. The actual computation steps were as follows:

First Iteration:

Step 1. Draw 500 samples to form S_{n0} randomly generating the $n_{i.}^*$ independently for each sample.

Step 2. The objective value of (14) is 0.45088.

Second Iteration:

Step 1. Draw 500 samples to add to S_{n0}.

Step 2. The objective value of (14) is 0. The final sampling design is attained and achieved the minimum value $E[w(s)]$5.0418.

This procedure took approximately 30 seconds using a Fortran program on a Sun Ultra 10 workstation.

Table 6
Expected Sample Cell Counts Under Proportionate Stratification (n = 151)

0.73	0.58	0.08	0.59	0.69	0.84	0.04	0.17	0.27	0.8	0.02	0.84	0.79	0.03	0.53	7
0.43	0.39	0.35	0.57	0.35	0.38	0.47	0.53	0.39	0.96	0.52	0.27	0.68	0.40	0.31	7
0.73	0.25	0.15	0.73	0.48	0.32	0.91	0.49	0.03	0.61	0.14	0.61	0.73	0.25	0.87	7
0.13	0.28	0.35	0.60	0.26	0.38	0.37	0.39	0.71	0.01	0.93	0.72	0.30	0.66	0.91	7
0.32	0.06	0.86	0.47	0.80	0.93	0.96	0.30	0.65	0.72	0.67	0.54	0.51	0.77	0.44	9
0.12	0.78	0.81	0.34	0.28	0.02	0.89	0.41	0.94	0.82	0.37	0.81	0.85	0.51	0.05	8
0.48	0.51	0.50	0.62	0.35	0.11	0.85	0.78	0.29	0.39	0.69	0.07	0.67	0.78	0.91	8
0.86	0.41	0.11	0.17	0.75	0.89	0.48	0.48	0.91	0.20	0.53	0.67	0.34	0.19	0.01	7
0.81	0.00	0.13	0.93	0.36	0.12	0.19	0.86	0.33	0.04	0.79	0.69	0.56	0.37	0.82	7
0.82	0.22	0.54	0.82	0.61	0.46	0.74	0.33	0.24	0.53	0.41	0.18	0.30	0.03	0.77	7
0.95	0.60	0.35	0.33	0.95	0.43	0.06	0.63	0.71	0.02	0.55	0.23	0.87	0.21	0.11	7
0.96	0.65	0.96	0.83	0.41	0.58	0.49	0.27	0.74	0.88	0.93	0.46	0.6	0.13	0.11	9
0.83	0.54	0.05	0.96	0.79	0.70	0.33	0.81	0.86	0.45	0.45	0.84	0.29	0.30	0.80	9
0.75	0.65	0.63	0.04	0.32	0.36	0.38	0.80	0.50	0.23	0.37	0.23	0.85	0.69	0.20	7
0.79	0.31	0.55	0.26	0.04	0.05	0.91	0.11	0.43	0.79	0.14	0.64	0.44	0.48	0.06	6
0.23	0.92	0.81	0.42	0.49	0.10	0.74	0.56	0.24	0.47	0.34	0.57	0.60	0.56	0.95	8
0.13	0.77	0.65	0.66	0.05	0.23	0.58	0.74	0.19	0.94	0.26	0.75	0.16	0.71	0.18	7
0.31	0.01	0.60	0.38	0.01	0.55	0.70	0.72	0.20	0.87	0.55	0.82	0.77	0.44	0.07	7
0.63	0.67	0.21	0.02	0.16	0.68	0.14	0.17	0.95	0.78	0.58	0.55	0.94	0.96	0.56	8
0.99	0.40	0.31	0.26	0.85	0.87	0.77	0.75	0.42	0.49	0.76	0.51	0.75	0.53	0.34	9
12	9	9	10	9	9	11	10	10	11	10	11	12	9	9	151

Table 7
Occupational Health Survey, Vihma (1981) Expected Sample Cell Counts Under Proportionate Stratification (n = 100)

Type of Industry	Number of Personnel			
	5–9	10–19	20–49	$r_{i.}$
Food products	2.38	3.56	3.78	9.72
Food	0.35	0.14	0.56	1.05
Beverage	0.14	0.07	0.21	0.42
Textiles	1.33	1.26	1.46	4.05
Apparel	3.15	3.71	2.09	8.95
Leather	0.56	0.14	0.07	0.77
Footwear	0.07	0.07	0.21	0.35
Wood Products	2.37	1.89	0.91	5.17
Furniture	1.33	0.84	0.91	3.08
Paper Products	0.42	0.49	0.42	1.33
Printing	7.20	6.01	4.20	17.41
Industrial Chemicals	0.56	0.35	0.28	1.19
Chemical Products	1.82	1.54	1.53	4.89
Petrolium	0.14	0.07	0.00	0.21
Misc Coal and Petrol.	0.07	0.07	0.14	0.28
Rubber Products	0.14	0.21	0.07	0.42
Plastic Products	1.40	1.05	1.19	3.64
Glass Products	0.42	0.21	0.21	0.84
Non-Metal Minerals	1.12	0.98	0.84	2.94
Iron & Steel	0.14	0.07	0.35	0.56
Nonferrous Metal	0.35	0.14	0.28	0.77
Fabricated Metal	4.96	4.06	2.59	11.61
Machinery	2.80	1.96	3.21	7.97
Electrical	1.89	1.60	1.33	4.82
Transport Equipment	0.84	0.84	0.84	2.52
Scientific Equipment	0.56	0.42	0.49	1.47
Manufacturing Industries	1.68	0.91	0.98	3.57
$n_{.j}$	38.19	32.66	29.15	100.00

4. Concluding Remarks

We propose a method for two-way stratification which extends the applicability of the linear programming approach of Sitter and Skinner (1994) to much larger problems. The method focuses on how to construct a small "representative" candidate sample set by using an unequal probability sampling procedure which generates candidate samples which nearly meet the AEPA constraints of the linear programming problem and then applying the linear programming method to this much smaller set.

It should be noted that the linear programming method extends easily to stratified multi-stage designs. Since there is no fundamental difference between the original linear programming approach and the extension proposed here, this is still true of the proposed method. In the same spirit, one can view discussion on issues around variance estimation of the resulting estimators in Sitter and Skinner (1994) as well.

One should also note that once one restricts to bracketing integers around the nP_{ij}'s, the problem is related to a controlled rounding problem (see Kelly, Golden and Assad 1993, and references therein), though we do not explore this aspect here.

Acknowledgements

This research was supported by a grant from the Natural Science and Engineering Research Council of Canada.

Appendix 1

Proof of (16): $n_{i.}(s) - [r_{i.}] \sim$ Bernoulli $(r_{i.} - \lfloor r_{i.} \rfloor)$ and has variance $(r_{i.} - \lfloor r_{i.} \rfloor)(1 + \lfloor r_{i.} \rfloor - r_{i.})$. This implies

$$\begin{aligned}\sum_s \left(n_{i.}(s) - r_{i.}\right)^2 p(s) &= E\left(n_{i.}(s) - r_{i.}\right)^2 V\left(n_{i.}(s)\right) \\ &= V\left(n_{i.}(s) - \lfloor r_{i.} \rfloor\right) \\ &= \left(r_{i.} - \lfloor r_{i.} \rfloor\right)\left(1 + \lfloor r_{i.} \rfloor - r_{i.}\right),\end{aligned}$$

and by similar argument that

$$\sum_s \left(n_{.j}(s) - r_{.j}\right)^2 p(s) = \left(r_{.j} - \lfloor r_{.j} \rfloor\right)\left(1 + \lfloor r_{.j} \rfloor - r_{.j}\right).$$

Therefore, with $w(s)$ defined in (3),

$$\begin{aligned}&E[w(s)] \\ &= \sum_s w(s)p(s) = \sum_s \left\{\sum_i \left(n_{i.}(s) - r_{i.}\right)^2 + \sum_j \left(n_{.j}(s) - r_{.j}\right)^2\right\} p(s) \\ &= \sum_i \sum_s \left(n_{i.}(s) - r_{i.}\right)^2 p(s) + \sum_j \sum_s \left(n_{.j}(s) - r_{.j}\right)^2 p(s) \\ &= \sum_i \left(r_{i.} - \lfloor r_{i.} \rfloor\right)\left(1 + \lfloor r_{i.} \rfloor - r_{i.}\right) + \sum_j \left(r_{.j} - \lfloor r_{.j} \rfloor\right)\left(1 + \lfloor r_{.j} \rfloor - r_{.j}\right).\end{aligned}$$

References

Bryant, E.C., Hartley, H.O. and Jessen, R.J. (1960). Design and estimation in two-way stratification. *Journal of the American Statistical Association*, 55, 105-124.

Causey, B.D., Cox, L.H. and Ernst, L.R. (1985). Applications of transportation theory to statistical problems. *Journal of the American Statistical Association*, 80, 903-909.

Engle, M., Marsden, G. and Pollock, S.W. (1971). Child work and social class. *Psychiatry*, 34, 140-150.

Goodman, R., and Kish, L. (1950). Controlled selection-a technique in probability sampling. *Journal of the American Statistical Association*, 45, 350-372.

Hess, I., Riedel, D.C. and Fitzpatrick, T.B. (1976). *Probability Sampling of Hospitals and Patients*. University of Michigan, Ann Arbor, second edition.

Jessen, R.J. (1970). Probability sampling with marginal constraints. *Journal of the American Statistical Association*, 65, 776-795.

Jessen, R.J. (1973). Some properties of probability lattice sampling. *Journal of the American Statistical Association*, 68, 20-28.

Jessen, R.J. (1975). Square and cubic lattice sampling. *Biometrics*, 31, 449-471.

Kelly, J.K., Golden, B.L. and Assad, A.A. (1993). The controlled rounding problem: complexity and computational experience. *European Journal of Operational Research*, 65, 207-217.

Lahiri, P., and Mukerjee, R. (2000). On a simplification of the linear programming approach to controlled sampling. *Statistical Sinica*, 10, 1171-1178.

Lu, W. (2000). Multi-way stratification by linear programming made practical. M.Sc. Thesis, Simon Fraser University.

Rao, J.N.K., Hartley, H.O. and Cochran, W.G. (1962). On a simple procedure of unequal probability sampling without replacement. *Journal of the Royal Statistical Society*, *Serie B*, 24, 482-491.

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

Rao, J.N.K., and Nigam, A.K. (1992). 'Optimal' controlled sampling: a unified approach. *International Statistical Review*, 60, 89-98.

Sitter, R.R., and Skinner, C.J. (1994). Multi-way stratification by linear programming. *Survey Methodology*, 20, 65-73.

Skinner, C.J., Holmes, D.J. and Holt, D. (1994). Multiple frame sampling for multiple stratification. *International Statistical Review*, 62, 333-347.

Vihma, T. (1981). Health hazards and stress factors in small industry-Prevalence study in the province of Uusimaa with special reference to the type of industry and the occupational title as classifications for the description of occupational health problems. *Scandinavian Journal of Work, Environment and Health*, 7, Suppl. 3, 1-149.

Waterton, J.J. (1983). An exercise in controlled selection. *Applied Statistics*, 32, 150-164.