MAXIMIZING AND MINIMIZING OVERLAP WHEN SELECTING ANY NUMBER OF UNITS PER STRATUM SIMULTANEOUSLY FOR TWO DESIGNS WITH DIFFERENT STRATIFICATIONS

Lawrence R. Ernst and Steven P. Paben, Bureau of Labor Statistics
Lawrence R. Ernst, BLS, 2 Massachusetts Ave., NE, Room 3160, Washington DC 20212

Key Words: Stratified designs, Transportation problems, Optimal

1. Introduction

Consider the following sampling problem. Sample units are to be selected simultaneously for two designs, denoted as D_1 and D_2, with generally different stratifications. Typically, the universes for the two designs are identical, although this is not assumed. The selection of sample units for each design is to be without replacement, with probability proportional to a measure of size (pps) that is generally different for the two designs. We wish to maximize the overlap of the sample units, that is to select units so that:

There are a predetermined number of units, n_{i1}, selected from D_1 stratum i and a predetermined number of units, n_{j2}, selected from D_2 stratum j; that is, the sample size for each stratum and design combination is fixed. (1.1)

Each unit in the universe is selected into each sample with its assigned probability. (1.2)

The expected number of units in common to the two samples is maximized. (1.3)

In this article we demonstrate how a variation of the two-dimension controlled selection procedure of Causey, Cox, and Ernst (1985) can be used to obtain samples that satisfy these conditions and, with a slight modification, the conditions for the analogous problem of minimizing the overlap of the sample units.

Many procedures have been developed for maximizing and minimizing the overlap of sample units since Keyfitz's (1951) pioneering work. Ernst (1999) discusses the various overlap procedures. The majority of these procedures have been developed for the following somewhat different application. Units are selected pps, without replacement for a survey with a stratified design. Later a new sample is to be selected using a new a size measure and a different stratification. To reduce costs it may be desirable to maximize the expected number of units common to the two samples while preserving prespecified selection probabilities in the new design. Minimization of overlap, in contrast, is typically employed as a method of reducing respondent burden.

In the redesign illustration just described, unlike the case for the present problem, the two samples must be selected sequentially, since the designs are for different points in time. Various procedures for overlap maximization and minimization of two samples when the samples are selected sequentially have been developed, but most have one or more of the following drawbacks: applicable to only one sample unit per stratum designs; requires that the stratifications for the two designs overlapped are identical; fails to generally attain the optimal overlap. The overlap procedure of Causey, Cox, and Ernst (1985), which formulates the problem as a transportation problem, has none of these drawbacks, but the size of the transportation problems are commonly so large that it is operationally infeasible to implement.

In contrast to the redesign illustration, there are other applications for which samples are selected at the same point in time for two or more surveys. Some overlap procedures have been developed specifically to be used for simultaneous selection and generally produce a better overlap than procedures developed for sequential selection or are computationally more efficient.

Ernst (1996, 1998) developed optimal, simultaneous procedures for two different situations. Ernst (1996) is only applicable to one unit per stratum designs, but the designs may have different stratifications. In Ernst (1998) there are no restrictions on the number of sample units per stratum, but the stratifications must be identical. Both procedures are applicable to both the maximization and minimization problems, but are restricted to the overlap of two designs. These two procedures employ the algorithm in Causey, Cox, and Ernst (1985) for solving the two-dimensional version of the controlled selection problem developed by Goodman and Kish (1950). This algorithm involves solving a sequence of transportation problems.

The procedure presented in this paper combines the features of the Ernst (1996) and Ernst (1998) procedures; that is, the procedure is an optimal, simultaneous procedure that has no restrictions on the number of units per stratum and is applicable when the two designs have different stratifications. The solution, although borrowing ideas from both of the earlier papers, is mostly a generalization of the Ernst (1996) procedure. However, it is substantially more complex than that procedure. In order to understand the need for this extra complexity, we present in Section 2 an outline of the direct generalization of the Ernst (1996) procedure for the maximization problem to other than one unit per stratum designs and demonstrate why this direct generalization can result in three problems that prevent it from producing a solution without

modifications. In Section 3 we present the main procedure for the maximization problem and explain how the modifications of the Ernst (1996) procedure overcome the three problems of Section 2; the proof of some of the claims in Section 3 are deferred to the Appendix, Section 6. Like both the Ernst (1996) and (1998) procedures, this new procedure requires the solution of a sequence of transportation problems. In Section 4 we show how to modify the procedure to solve the minimization problem. In Section 5 we report the results of a simulation study that illustrates a potential application of the new procedure. Also included in this section is a discussion of operational upper limits on the size of the universe to which this procedure can be applied.

Due to space limitations, the complete paper is not presented here. Omissions include all of Sections 4-6, the list of references, and most of the tables and figures. The complete paper is available from the authors.

2. Problems with Directly Generalizing the Ernst (1996) Procedure

In this section we will: introduce some notation; reformulate (1.2) and (1.3) in terms of the notation; illustrate by means of an example the direct generalization of Ernst (1996) to cases where at least one of the designs is not one unit per stratum; and use this example to demonstrate the three problems with this direct generalization that require the modifications presented in the next section.

Let M, N denote the number of D_1 and D_2 strata, respectively. If the universes for the two designs are not identical then we artificially create identical universes as follows. If a unit is in D_1 only, arbitrarily assign it to some D_2 stratum and set its D_2 selection probability to 0. Units in D_2 only are treated analogously.

For $i = 1,...,M$, $j = 1,...,N$, let D_{i1}, D_{j2} denote the set of units in D_1 stratum i and D_2 stratum j, respectively; let $D_{ij}^* = D_{i1} \cap D_{j2}$ and let t_{ij} denote the number of units in D_{ij}^*. We denote the set of all units in the two designs by the set of ordered triples $T = \{(i,j,k) : i = 1,...,M,\ j = 1,...,N,\ k = 1,...,t_{ij}\}$. Let π_{ijk1}, π_{ijk2} denote the D_1, D_2 selection probability, respectively, for $(i,j,k) \in T$. Let

$$\pi'_{ijk3} = \min\{\pi_{ijk1}, \pi_{ijk2}\},\ \pi'_{ijk\alpha} = \pi_{ijk\alpha} - \pi'_{ijk3},\ \alpha = 1,2,$$

$$\text{and } \pi'_{ijk4} = 1 - \sum_{\alpha=1}^{3} \pi'_{ijk\alpha} \qquad (2.1)$$

Let S_1, S_2 denote the random sets consisting of the sample units for D_1, D_2, respectively. Let S'_1, S'_2, S'_3, S'_4 be the random sets denoting the set of units, respectively: in S_1 but not in S_2, in S_2 but not in S_1, in both samples, and in neither sample.

In terms of our notation (1.2) and (1.3) are equivalent to, respectively,

$$\Pr((i,j,k) \in S_\alpha) = \pi_{ijk\alpha},\ (i,j,k) \in T,\ \alpha = 1,2 \qquad (2.2)$$

$$\Pr((i,j,k) \in S'_3) \text{ is maximal for each } (i,j,k) \in T \qquad (2.3)$$

To establish (2.2) and (2.3) it is sufficient to show that

$$\Pr((i,j,k) \in S'_\beta = \pi'_{ijk\beta}),\ (i,j,k) \in T,\ \beta = 1,2,3 \qquad (2.4)$$

since (2.1) and (2.4) imply (2.2), while (2.4) with $\beta = 3$, together with fact that $\Pr((i,j,k) \in S'_3) \le \pi'_{ijk3}$, $(i,j,k) \in T$, for any selection procedure satisfying (2.2), imply (2.3).

We use the following example to illustrate the direct generalization of the Ernst (1996) procedure and to explain the three reasons that this generalization does not work without modifications unless both designs are one unit per stratum. In this example: $M = 3$, $N = 2$; $n_{i1} = 1$, $n_{j2} = 2$ for all i,j; the two designs have the same eight units, with $t_{11} = t_{22} = 2$, $t_{ij} = 1$ for all other i,j; and the selection probabilities for the eight units are given in Table 1 at the end of the paper.

In general we begin the direct generalization of the Ernst (1996) procedure by constructing an $(M+2)\times(N+2)$ array, $\mathbf{A} = (a_{ij})$, of expected values. For $i = 1,...,M$, $j = 1,...,N$, the expected number of units in $D_{ij}^* \cap S'_3$ is a_{ij}; the expected number of units in $D_{i1} \cap S'_1$ is $a_{i(N+1)}$; and the expected number of units in $D_{j2} \cap S'_2$ is $a_{(M+1)j}$. Then, in order to satisfy (2.4), we must have

$$a_{ij} = \sum_{k=1}^{t_{ij}} \pi'_{ijk3},\quad a_{i(N+1)} = \sum_{j=1}^{N}\sum_{k=1}^{t_{ij}} \pi'_{ijk1}$$

$$a_{(M+1)j} = \sum_{i=1}^{M}\sum_{k=1}^{t_{ij}} \pi'_{ijk2},\quad i = 1,...,M,\ j = 1,...,N$$

Furthermore, $a_{(M+1)(N+1)} = 0$ and the remaining cells are marginals. (We refer to an array, such as **A**, in which the final row and final column are marginal values as a tabular array.) **A** for the example is presented below

$$\mathbf{A} = \begin{array}{cccc} .6 & .4 & 0 & 1 \\ .4 & .6 & 0 & 1 \\ .2 & .6 & .2 & 1 \\ .8 & .4 & 0 & 1.2 \\ \hline 2 & 2 & .2 & 4.2 \end{array} \qquad (2.5)$$

The next step in the procedure is to obtain a solution to the controlled selection problem corresponding to **A** using the procedure of Causey, Cox, and Ernst (1985). A controlled rounding of a real-valued, tabular array **A** is an integer-valued, tabular

array $\mathbf{M}$ with the same dimensions as $\mathbf{A}$ that rounds every element a_{ij} of $\mathbf{A}$ that is not an integer to either the next integer above or the next integer below a_{ij} and leaves integer elements of $\mathbf{A}$ unchanged. For example, the three arrays $\mathbf{M}_1, \mathbf{M}_2, \mathbf{M}_3$ below are all controlled roundings of $\mathbf{A}$. Cox and Ernst (1982) demonstrated that a controlled rounding of a tabular array always exists and can be obtained by modeling the controlled rounding problem as a transportation problem. A set, $\mathbf{M}_1 = (m_{ij1})$, $\mathbf{M}_2 = (m_{ij2})$,..., $\mathbf{M}_\ell = (m_{ij\ell})$, of controlled roundings of $\mathbf{A}$, and associated probabilities, $p_1,...,p_\ell$, satisfying

$$\sum_{u=1}^{\ell} m_{iju} p_u = a_{ij}, \quad i = 1,...,M+2, \quad j = 1,...,N+2 \qquad (2.6)$$

is known as a solution to the controlled selection problem $\mathbf{A}$. For example, the arrays

$$\mathbf{M}_1 = \begin{array}{cccc} 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ \hline 2 & 2 & 0 & 4 \end{array} \quad \mathbf{M}_2 = \begin{array}{cccc} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ \hline 2 & 2 & 0 & 4 \end{array} \quad \mathbf{M}_3 = \begin{array}{cccc} 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 2 \\ \hline 2 & 2 & 1 & 5 \end{array}$$

with associated probabilities .2,.6,.2, respectively, is a solution to (2.1).

A single array $\mathbf{M}_u$ is selected from among $\mathbf{M}_1,...,\mathbf{M}_\ell$ using the associated probabilities. Then for $i = 1,..,M$, $j = 1,...,N$: m_{iju} is the number of units in D_{ij}^* to be selected to be in S_3', with the selection among these t_{ij} units proportional to π'_{ijk3}; $m_{i(N+1)u}$ is the number of units in D_{i1} to be selected to be in S_1', with the selection with probability proportional to π'_{ijk1}; and $m_{(M+1)ju}$ is the number of units in D_{j2} to be selected to be in S_2', with the selection with probability proportional to π'_{ijk2}. For this example, three problems arise in the selection process because the D_2 stratification is not 1 unit per stratum and hence the first two column totals are not 1.

To illustrate the first problem for this example, assume that $\mathbf{M}_1$ has been selected. Then $(1,2,1) \in S_3'$ since $m_{121} = 1$, but $(1,2,1) \in S_2'$ also because $m_{124} = 1$, a contradiction.

To illustrate the second problem, again assume that $\mathbf{M}_1$ has been selected. Then $(1,1,1) \notin S_3'$ since $m_{111} = 0$, while $(1,1,1) \notin S_2'$ either since $m_{141} = 0$. Consequently, $(1,1,1) \notin S_2$ if $\mathbf{M}_1$ is selected and (2.2) cannot be satisfied since $\pi_{1112} = 1$.

To illustrate the third problem, assume that $\mathbf{M}_2$ has been selected. Then one of the units $(1,1,1),(1,1,2)$ would be selected to be in S_3' since $m_{112} = 1$, while one of the two units would be selected to be in S_2' since $m_{142} = 1$. The Ernst (1996) procedure selects the units corresponding to each cell independently. This will not work here since a unit could be selected to be in both S_3' and S_2'.

3. The Main Procedure

We divide this section into three subsections as follows.

In Section 3.1, given a set of probabilities $\pi'_{ijk\beta}$, $(i,j,k) \in T$, $\beta = 1,2,3,4$, we construct an array $\mathbf{A}$ of expected values. This is analogous to the array $\mathbf{A}$ of Section 2, but more complex in order to avoid problems 1 and 2 of Section 2. We also obtain a controlled rounding $\mathbf{M}$ of $\mathbf{A}$, which determines the actual number of sample units to be in S_1', S_2', S_3', S_4' by type of unit.

In Section 3.2 we describe how to select a single sample for the two designs given $\mathbf{M}$. By a sample, we mean the following. Each unit in T must be in exactly one of the four sets S_1', S_2', S_3', S_4'. A sample simply specifies to which one of these four sets each unit in T belongs. The approach of associating a single sample with each controlled rounding $\mathbf{M}$ differs from the approach in Ernst (1996) where, as described in Section 2 of the current paper, each controlled rounding is used together with a probability mechanism to select a sample. The approach is different here to avoid the third problem of Section 2.

The algorithm described in Sections 3.1 and 3.2 result in a single sample. However, what we need is a set of samples, $S_{1u}', S_{2u}', S_{3u}', S_{4u}'$, $u = 1,...,\ell$, and associated probabilities, $p_1,...,p_\ell$, where: ℓ is the number of samples; S_{1u}' is the set of units in the D_1 sample only for sample u, with analogous definitions for $S_{2u}', S_{3u}', S_{4u}'$; and p_u is the probability of selecting sample u. Note that for each u, each unit in T is in exactly one of $S_{1u}', S_{2u}', S_{3u}', S_{4u}'$. To illustrate, for the example in Section 2, a possible set of samples is given in Table 2 of the full paper. Here $\ell = 4$ and the probabilities associated with the four samples are .4,.2,.2,.2, respectively. The construction of the ℓ samples is described in Section 3.3 and employs a recursive procedure that requires the construction of an array of expected values $\mathbf{A}_u$ and a controlled rounding $\mathbf{M}_u$ of $\mathbf{A}_u$ for each sample u.

3.1. The Construction of $\mathbf{A}$

To construct an array $\mathbf{A}$ that overcomes the first two problems of Section 2 we begin by partitioning T into five different sets, namely:

$$T_{1C} = \{(i,j,k) : \pi_{ijk2} < \pi_{ijk1} = 1\}$$

$$T_{1S} = \{(i,j,k) : \pi_{ijk2} < \pi_{ijk1} < 1\}$$

$$T_{2C} = \{(i,j,k) : \pi_{ijk1} < \pi_{ijk2} = 1\}$$
$$T_{2S} = \{(i,j,k) : \pi_{ijk1} < \pi_{ijk2} < 1\}$$
$$T_3 = \{(i,j,k) : \pi_{ijk1} = \pi_{ijk2}\}$$

For the example of Section 2, $T_{1C} = \varnothing$, $T_{1S} = \{(3,1,1)\}$, $T_{2C} = \{(1,1,1)\}$, $T_{2S} = \{(1,1,2),(1,2,1)\}$, $T_3 = \{(2,1,1),(2,2,1),(2,2,2),(3,2,1)\}$.

As we will show, the partitioning by the numerical subscript overcomes the first problem of Section 2, the further partitioning determined by C and S overcomes the second problem. We will accomplish this by using an expanded tabular array $\mathbf{A} = (a_{ij})$ with dimensions $M^* \times N^*$, where $M^* = 3M + N + 2$ and $N^* = M + 3N + 2$, instead of the array (2.5) of dimensions $(M+2)\times(N+2)$ described in Section 2. The expanded **A** contains five subarrays corresponding to the five sets in the partition. The subarray corresponding to T_{1C} is denoted by $\mathbf{A}_{1C}$ with analogous notation for the other four subarrays. Each subarray corresponds to the internal elements in (2.5), except that each subarray is restricted to the units in the corresponding subset. Furthermore, instead of dimensions $(M+1)\times(N+1)$, $\mathbf{A}_{1C}, \mathbf{A}_{1S}$ have dimensions $M \times (N+1)$; $\mathbf{A}_{2C}, \mathbf{A}_{2S}$ have dimensions $(M+1)\times N$; and $\mathbf{A}_3$ has dimensions $M \times N$. This is because units in T_{1C}, T_{1S} cannot be in S_2'; units in T_{2C}, T_{2S} cannot be in S_1'; and units in T_3 cannot be in either S_1' or S_2'. These five subarrays allows us to separately control the number of units selected of each of these five types, which is the key to overcoming the first two problems of Section 2.

We proceed to define these five subarrays. **A** for the example is presented at the end of this subsection, with the boundaries of the five subarrays indicated by broken lines. In this figure, the first row and first column are not elements of **A**, but instead list the row and column numbers, respectively, of **A**, with those row and columns consisting entirely of zeros omitted to conserve space. (As a result, the broken line below row 7, across the first three columns is omitted.) Let $T_{ij1C} = \{k : (i,j,k) \in T_{1C}\}$, $i = 1,\ldots,M$, $j = 1,\ldots,N$, with analogous definitions for $T_{ij1S}, T_{ij2C}, T_{ij2S}, T_{ij3}$. $\mathbf{A}_3$ occupies the upper left-hand corner of **A** with its elements defined by

$$a_{ij} = \sum_{k \in T_{ij3}} \pi'_{ijk3}, \quad i = 1,\ldots,M, \quad j = 1,\ldots,N \qquad (3.1)$$

$\mathbf{A}_{2C}$ is located to the right of $\mathbf{A}_3$ and $\mathbf{A}_{2S}$ to the right of $\mathbf{A}_{2C}$. Similarly, $\mathbf{A}_{1C}$ is located below $\mathbf{A}_3$ and $\mathbf{A}_{1S}$ below $\mathbf{A}_{1C}$. $\mathbf{A}_{2C}$ begins in column $N+2$, not $N+1$, and $\mathbf{A}_{1C}$ begins in row $M+2$, so that these two subarrays do not overlap. Consequently, the cells in the first $M+1$ rows of column $N+1$ of **A** and the cells in the first $N+1$ columns of row $M+1$ are not in any of the five subarrays and we let $a_{ij} = 0$ for each of these cells. An essential reason for the placement of the five subarrays as described, which we will discuss further later, is to insure that none of the other subarrays have cells in the same columns as $\mathbf{A}_{2C}, \mathbf{A}_{2S}$ or the same rows as $\mathbf{A}_{1C}, \mathbf{A}_{1S}$.

The first M rows of $\mathbf{A}_{2C}, \mathbf{A}_{2S}$ are defined as in (3.1), except j is replaced by $j+N+1$ and $j+2N+1$ for $\mathbf{A}_{2C}$ and $\mathbf{A}_{2S}$, respectively, on the left-hand side of (3.1) only; while T_{ij3} is replaced by T_{ij2C} and T_{ij2S}, respectively. The cells in the first N columns of $\mathbf{A}_{1C}, \mathbf{A}_{1S}$ are defined by making analogous substitutions in (3.1). As for row $M+1$ in $\mathbf{A}_{2C}$, the row to be used in selecting units in S_2', we let

$$a_{(M+1)(j+N+1)} = \sum_{i=1}^{M} \sum_{k \in T_{ij2C}} \pi'_{ijk2}, \quad j = 1,\ldots,N \qquad (3.2)$$

while the same formula holds for row $M+1$ of $\mathbf{A}_{2S}$, except N is replaced by $2N$ on the left-hand side and C is replaced by S on the right-hand side. For column $N+1$ of $\mathbf{A}_{1C}$ we analogously have

$$a_{(i+M+1)(N+1)} = \sum_{j=1}^{N} \sum_{k \in T_{ij1C}} \pi'_{ijk1}, \quad i = 1,\ldots,M \qquad (3.3)$$

while for column $N+1$ of $\mathbf{A}_{1S}$ we replace M by $2M$ on the left-hand side of (3.3) and C by S on the right-hand side.

We let $a_{ij} = 0$ for the remaining elements in the first $3M+1$ rows and first $3N+1$ columns of **A**. Cells defined to be 0 have no role in the sample selection process. We postpone the definition of the cells that are in either the final N internal rows or the final M internal columns of **A**. For the example we have so far defined elements that are in both the first 10 rows and first 7 columns of **A**.

Corresponding to **A**, we obtain a controlled rounding of this array, **M**, which is used to select the first sample. We first explain the meaning of those elements of **A** that are within the five subarrays and the corresponding elements of **M**. $a_{i(j+N+1)}$, $i = 1,\ldots,M$, $j = 1,\ldots,N$, the value for cell (i,j) of $\mathbf{A}_{2C}$ is the expected number of units in $D_{ij}^* \cap T_{2C}$ to be selected to be in S_3' and $m_{i(j+N+1)}$ is the actual number of such units to be selected for the first sample. Likewise, $a_{(M+1)(j+N+1)}$ is the expected number of units in $D_{j2} \cap T_{2C}$ to be selected to be in S_2' and

$m_{(M+1)(j+N+1)}$ is the actual number of such units for the first sample. The cell values for the other four arrays have analogous interpretations.

We now explain the need for the final N internal rows in **A**. Let

$$a'_{j2} = \sum_{i=1}^{3M+1} a_{ij},\ m'_{j2} = \sum_{i=1}^{3M+1} m_{ij},\ j = 1,...,3N+1 \qquad (3.4)$$

$$a''_{j2} = a'_{j2} + a'_{(j+N+1)2} + a'_{(j+2N+1)2},$$

$$m''_{j2} = m'_{j2} + m'_{(j+N+1)2} + m'_{(j+2N+1)2},\ j = 1,...,N \qquad (3.5)$$

Then from (3.4), (3.5), and the discussion in the previous paragraph, it follows that the three terms in the definition of a''_{j2} are the expected number of units in $D_{j2} \cap (T_3 \cup T_{1C} \cup T_{1S})$, $D_{j2} \cap T_{2C}$, $D_{j2} \cap T_{2S}$, respectively, to be selected to be in S_2; consequently, a''_{j2} is the expected number of units in D_{j2} to be selected to be in S_2. From (2.1), (3.1), (3.2), (3.4) it follows that for $j = 1,...,N$,

$$a'_{j2} = \sum_{i=1}^{M} \sum_{k \in T_{ij3} \cup T_{ij1C} \cup T_{ij1S}} \pi_{ijk2},\ a'_{(j+N+1)2} = \sum_{i=1}^{M} \sum_{k \in T_{ij2C}} \pi_{ijk2},$$

$$a'_{(j+2N+1)2} = \sum_{i=1}^{M} \sum_{k \in T_{ij2S}} \pi_{ijk2} \qquad (3.6)$$

and, consequently, that $a''_{j2} = n_{j2}$ as required by (1.1). Furthermore, since m''_{j2} is the actual number of units in D_{j2} to be selected to be in S_2 for the first possible sample and since $a''_{j2} = n_{j2}$, we must also have $m''_{j2} = a''_{j2}$. To force this last relationship to be true for any controlled rounding **M** of **A** we define elements in the last N internal rows of **A** as follows. For any real number x, let $\lceil x \rceil$ be the smallest integer that is greater than or equal to x and let $r(x) = \lceil x \rceil - x$. Then let

$$a_{(j+3M+1)j} = r(a'_{j2}),$$

$$a_{(j+3M+1)(j+2N+1)} = r(a'_{(j+2N+1)2}),\ j = 1,...,N \qquad (3.7)$$

and let the cell value be 0 for all other internal cells in row $j+3M+1$ of **A**. It is established in the Appendix that $m''_{j2} = a''_{j2}$ and illustrated in the full paper why (3.7) is needed to force $m''_{j2} = a''_{j2}$.

The entries in the final M internal columns of **A** are defined analogously to entries in the final N internal columns, that is

$$a'_{i1} = \sum_{j=1}^{3N+1} a_{ij},\ i = 1,...,3M+1 \qquad (3.8)$$

$$a_{i(i+3N+1)} = r(a'_{i1}),$$

$$a_{(i+2M+1)(i+3N+1)} = r(a'_{(i+2M+1)1}),\ i = 1,...,M \qquad (3.9)$$

and the cell value is 0 for all other internal elements in column $i+3N+1$ of **A**. (3.8), (3.9) are needed to force the number of units in D_{i1} selected to be in S_1 for the first possible sample to be n_{i1}, $i = 1,...,M$.

This completes the definition of the internal elements of **A**. The remaining elements are the marginals. **M** is any controlled rounding of **A**

		1	2	3	4	6	7	10	11
	1	0	0	0	.4	.2	.4	0	1
	2	.4	.6	0	0	0	0	0	1
	3	0	.6	0	0	0	0	.4	1
A =	4	0	0	0	.6	.2	.4	0	1.2
	10	.2	0	.2	0	0	0	.6	1
	11	.4	0	0	0	.6	0	0	1
	12	0	.8	0	0	0	.2	0	1
	13	1	2	.2	1	1	1	1	7.2

3.2. Selection of a Sample Given **M**

We now describe how to select a single sample, that is a set of units in S'_1, S'_2, S'_3, S'_4 given **M**, which will be the first sample in the solution. For cell (i, j) in $\mathbf{A}_3$, select any m_{ij} units in $D^*_{ij} \cap T_3$ to be in S'_3, with the additional requirements that any unit (i, j, k) for which $\pi'_{ijk3} = 1$ must be selected and no unit for which $\pi'_{ijk3} = 0$ may be selected. Such a selection can always be made if there are at least m_{ij} units in $D^*_{ij} \cap T_3$ for which $\pi'_{ijk3} > 0$ and no more than m_{ij} units in $D^*_{ij} \cap T_3$ for which $\pi'_{ijk3} = 1$. It can be shown that the first of these conditions is met by combining (3.1) and $m_{ij} \le \lceil a_{ij} \rceil$, while the second condition follows from (3.1) and $\lfloor a_{ij} \rfloor \le m_{ij}$. We select the units in S'_3 from the corresponding cells of $\mathbf{A}_{1C}, \mathbf{A}_{1S}, \mathbf{A}_{2C}, \mathbf{A}_{2S}$ in the same way.

After all the units to be in S'_3 are selected, units are selected corresponding to the cells in the last row of $\mathbf{A}_{2C}, \mathbf{A}_{2S}$ to be in S'_2 as follows. For cell j of $\mathbf{A}_{2C}$ in this row, which is cell $(M+1, j+N+1)$ of **A**, choose any $m_{(M+1)(j+N+1)}$ units in $D_{j2} \cap T_{2C}$ to be in S'_2 among those units in $D_{j2} \cap T_{2C}$ not selected to be in S'_3. Units are selected similarly corresponding to cells in the last row of $\mathbf{A}_{2S}$. In the Appendix we show that this selection of units in $D_{j2} \cap T_{2C}$ and in $D_{j2} \cap T_{2S}$ to be in S'_2 avoids problems 1 and 2 in Section 2. It is here that we make use of the fact that that the columns containing elements of $\mathbf{A}_{2C}, \mathbf{A}_{2S}$ do not contain elements of any other subarray.

The selection of the units corresponding to the cells in the last column of $\mathbf{A}_{1C}, \mathbf{A}_{1S}$ to be in S'_1 is analogous to the selection of the units corresponding to the last row in $\mathbf{A}_{2C}, \mathbf{A}_{2S}$ to be in S'_2.

3.3. Recursively Selecting a Set of Samples

The selection of the ℓ samples $S'_{1u}, S'_{2u}, S'_{3u}, S'_{4u}$ and associated probabilities p_u, $u = 1,...,\ell$, is done recursively as follows. Sample 1 is simply the sample obtained in Section 3.2. To obtain sample u and p_u we begin with a set of probabilities $p_1,...,p_{u-1}$ for $u > 1$ and a set of probabilities $\pi'_{ijk\beta u}$, $(i,j,k) \in T$, $\beta = 1,2,3,4$, for $u \geq 1$, satisfying:

$$0 \leq \pi'_{ijk\beta u} \leq 1 \tag{3.10}$$

$$\sum_{\beta=1}^{4} \pi'_{ijk\beta u} = 1 \tag{3.11}$$

$$\text{If } \pi'_{ijk\beta} = 0 \text{ or } \pi'_{ijk\beta} = 1, \text{ then } \pi'_{ijk\beta u} = \pi'_{ijk\beta} \tag{3.12}$$

$$\sum_{i=1}^{M}\sum_{k=1}^{t_{ij}} \pi_{ijk2u} = n_{j2}, \quad j = 1,...,N, \text{ and}$$

$$\sum_{j=1}^{N}\sum_{k=1}^{t_{ij}} \pi_{ijk1u} = n_{i1}, \quad i = 1,...,M \tag{3.13}$$

where

$$\pi_{ijk\alpha u} = \pi'_{ijk\alpha u} + \pi'_{ijk3u}, \quad \alpha = 1,2 \tag{3.14}$$

For $u = 1$ we have $\pi'_{ijk\beta 1} = \pi'_{ijk\beta}$ for all i,j,k,β, and hence the set of $\pi'_{ijk\beta 1}$ satisfy (3.10)-(3.13). For general u, using $\pi'_{ijk\beta u}$ in place of $\pi'_{ijk\beta}$, an array $\mathbf{A}_u$ is constructed exactly as $\mathbf{A}$ was constructed in Section 3.1. In particular, for sample u, T_{1C} and the other four subsets that form a partition of T depend on π_{ijk1u}, π_{ijk2u}, not π_{ijk1}, π_{ijk2}. Therefore, a unit can be in different subsets for different u. Table 2 in the full paper lists the subset for each unit and sample for our example.

A controlled rounding $\mathbf{M}_u$ of $\mathbf{A}_u$ is then obtained and the sample $S'_{1u}, S'_{2u}, S'_{3u}, S'_{4u}$ selected exactly as sample 1 was selected in Section 3.2, except $\mathbf{M}_u$ replaces $\mathbf{M}$ and $\pi'_{ijk\beta u}$ replaces $\pi'_{ijk\beta}$. In particular, $\mathbf{A}_1 = \mathbf{A}$ and $\mathbf{M}_1 = \mathbf{M}$.

After sample u is selected we compute p_u as a function of sample u, the $\pi'_{ijk\beta u}$, and $p_1,...,p_{u-1}$, and then recursively compute $\pi'_{ijk\beta(u+1)}$ as follows. For $(i,j,k) \in T$, $\beta = 1,2,3,4$, let

$$\pi^*_{ijk\beta u} = \pi'_{ijk\beta u} \text{ if } (i,j,k) \in S'_{\beta u}, \ \pi^*_{ijk\beta u} = 1 - \pi'_{ijk\beta u} \text{ if } (i,j,k) \notin S'_{\beta u} \tag{3.15}$$

$$p^*_u = \min\{\pi^*_{ijk\beta u} : (i,j,k) \in T, \beta = 1,2,3,4\} \tag{3.16}$$

$$p_1 = p^*_1 \text{ if } u = 1, \quad p_u = \left(1 - \sum_{\gamma=1}^{u-1} p_\gamma\right) p^*_u \text{ if } u > 1 \tag{3.17}$$

$$\lambda_{ijk\beta u} = 1 \text{ if } (i,j,k) \in S'_{\beta u} \text{ and } \lambda_{ijk\beta u} = 0 \text{ if } (i,j,k) \notin S'_{\beta u} \tag{3.18}$$

and finally, if $p^*_u < 1$, let

$$\pi'_{ijk\beta(u+1)} = \frac{\pi'_{ijk\beta u} - \lambda_{ijk\beta u} p^*_u}{1 - p^*_u} \tag{3.19}$$

In the Appendix, we show that if (3.10)-(3.13) are satisfied for u then they are satisfied for $u+1$.

Finally, we need to explain how the recursive process terminates. Eventually, as is established in the Appendix,

there is an integer ℓ for which $\pi'_{ijk\beta\ell} = 1$ for exactly one β and $\pi'_{ijk\beta\ell} = 0$ for the other three β for each $(i,j,k) \in T$ (3.20)

Then there is only one possible sample ℓ, namely the sample for which

$$\lambda_{ijk\beta\ell} = \pi'_{ijk\beta\ell} \text{ for all } i,j,k,\beta \tag{3.21}$$

Then $p^*_\ell = 1$ by (3.15), (3.16); consequently,

$$p_\ell = 1 - \sum_{u=1}^{\ell-1} p_u \tag{3.22}$$

which ends the algorithm. It is established in the Appendix that this set of ℓ samples satisfies

$$\sum_{\gamma=1}^{\ell} \lambda_{ijk\beta\gamma} p_\gamma = \pi'_{ijk\beta}, \quad (i,j,k) \in T, \ \beta = 1,2,3,4 \tag{3.23}$$

which is equivalent to (2.4).

The results of using the recursive algorithm for the example are presented in the full paper. Here $\ell = 4$. Arrays $\mathbf{A}_u, \mathbf{M}_u$, $u = 1,2,3,4$, are presented, along with the samples in Table 3 and the $\pi'_{ijk\beta u}$ in Table 4. We also have that the p^*_u are .4,.33,.5,1, respectively, and that the p_u are .4,.2,.2,.2, respectively.

Any opinions expressed in this paper are those of the authors and do not constitute policy of the Bureau of Labor Statistics.

Table 1. Selection Probabilities for Example

	(i,j,k)							
	(1,1,1)	(1,1,2)	(1,2,1)	(2,1,1)	(2,2,1)	(2,2,2)	(3,1,1)	(3,2,1)
π_{ijk1}	.4	.2	.4	.4	.4	.2	.4	.6
π_{ijk2}	1.0	.4	.8	.4	.4	.2	.2	.6