Journal of Official Statistics, Vol. 18, No. 2, 2002, pp. 185–202

Maximizing and Minimizing Overlap When Selecting Any Number of Units per Stratum Simultaneously for Two Designs with Different Stratifications

Lawrence R. Ernst[1] *and Steven P. Paben*[1]

A number of procedures have been developed, beginning with the work of Keyfitz, for maximizing or minimizing the overlap of sampling units for two stratified designs. Certain overlap procedures have been developed for use when the two samples may be selected simultaneously. They generally produce a better overlap than procedures developed for sequential selection applications or are computationally more efficient. We present here a simultaneous overlap procedure developed from two previous overlap procedures of Ernst. One of these procedures is applicable when the stratifications for the two designs may be different, but is restricted to one unit per stratum designs. The other procedure has no restrictions on the number of sample units per stratum, but requires that the designs have identical stratifications. The new procedure does not have the restrictions of the previous procedures; that is, there are no restrictions on the number of sample units per stratum, nor is there a requirement that the two designs have identical stratifications. This procedure, like the two previous procedures, produces an optimal overlap and requires the solution of a sequence of transportation problems.

Key words: Stratified designs; transportation problems; optimal.

1. Introduction

Consider the following sampling problem. Sample units are to be selected simultaneously for designs for two different surveys, denoted as D_1 and D_2, for which the D_1 and D_2 stratifications are generally different. 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)

[1] U.S. Bureau of Labor Statistics, 2 Massachusetts Avenue, NE, Room 3160, Washington, DC 20212, U.S.A. Emails: Ernst_L@bls.gov and Paben_S@bls.gov

Acknowledgments: The authors thank the referee and Associate Editor for their valuable comments. Any opinions expressed in this article are those of the authors and do not constitute policy of the U.S. Bureau of Labor Statistics.

The expected number of units 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 minor modifications, 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. 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 size measure and generally 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 for the units in the new design, either selection probabilities for individual units or selection probabilities for the possible sets of sample units in a new stratum. For example, when the units being overlapped are primary sampling units (PSUs), which are geographic areas, an overlap maximization procedure can reduce the costs associated with hiring a new interviewer; when the units are ultimate sampling units, such a procedure can reduce the extra costs of an initiation interview. Minimization of overlap, in contrast, is typically employed as a method of reducing respondent burden.

In the redesign illustration just described, unlike the case of the present problem, the two samples must be selected sequentially, since the designs are for different points in time. In contrast, there are other applications for which samples are selected at the same point in time for two surveys with different measures of size and possibly different stratifications. Although overlap procedures developed for sequential selection can also be used in the case of simultaneous selection, 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 (1999) discusses the various overlap procedures of both types including their limitations.

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. These two procedures are applicable to both the maximization and minimization problems. Unlike these two procedures, previous simultaneous procedures either fail to attain the optimal overlap or do not guarantee a fixed sample size. Both 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 in this article 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 that it incorporates overcome the three problems of Section 2. The proofs of some of the claims in Section 3 are deferred until the Appendix, Section 6. Like both the Ernst (1996) and Ernst (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 the use of the new procedure.

One drawback to the procedure is that pairs of units, unlike individual units, are not selected into the D_1 and D_2 samples with preassigned probabilities. Furthermore, although joint selection probabilities can be calculated, they can be zero for some pairs of units, which would preclude the computation of unbiased variance estimates. The Ernst (1998) procedure has the same drawback and this issue is discussed there in further detail.

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 zero. Units in D_2 only are treated analogously.

For $i = 1, \ldots, M$, $j = 1, \ldots, N$, 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 $T = \{(i,j,k) : i = 1, \ldots, M,\ j = 1, \ldots, N,\ k = 1, \ldots, t_{ij}\}$. For $(i,j,k) \in T$, let π_{ijk1}, π_{ijk2} denote the D_1, D_2 selection probabilities, respectively, for (i,j,k) and let

$$\pi'_{ijk3} = \min\{\pi_{ijk1}, \pi_{ijk2}\}, \pi'_{ijk\alpha} = \pi_{ijk\alpha} - \pi'_{ijk3}, \alpha = 1,2, \text{and } \pi'_{ijk4} = 1 - \sum_{\alpha=1}^{3} \pi'_{ijk\alpha} \quad (2.1)$$

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 this notations (1.2) and (1.3) are equivalent to, respectively,

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

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

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

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

since (2.1) and (2.4) imply (2.2); while (2.4) with $\beta = 3$, together with the fact that $\Pr((i,j,k) \in S_3') \leq \pi'_{ijk3}$, $(i,j,k) \in T$, 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.

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

In general we would 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 $i = 1,\ldots,M$, $j = 1,\ldots,N$, a_{ij} is the expected number of units in $D^*_{ij} \cap S_3'$; $a_{i(N+1)}$ is the expected number of units in $D_{i1} \cap S_1'$; and $a_{(M+1)j}$ is the expected number of units in $D_{j2} \cap S_2'$. Then, in order to satisfy (2.4), we must have

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

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

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

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

The next step in the procedure is to obtain a solution to the controlled selection problem corresponding to $\mathbf{A}$ using the procedure of Causey, Cox, and Ernst (1985). A controlled rounding of a real-valued, tabular array $\mathbf{A}$ 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. (Our terminology differs here from that of Causey, Cox, and Ernst (1985), which refers to such roundings as zero-restricted controlled roundings.) 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, $\mathbf{M}_1 = (m_{ij1}), \mathbf{M}_2 = (m_{ij2}), \ldots, \mathbf{M}_\ell = (m_{ij\ell})$, of controlled roundings of $\mathbf{A}$ and associated probabilities of selection, $p_1, \ldots, p_\ell$, satisfying

$$\sum_{u=1}^{\ell} m_{iju} p_u = a_{ij},\ i = 1, \ldots, M+2,\ j = 1, \ldots, N+2 \tag{2.6}$$

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, constitute a solution to (2.5).

A single array $\mathbf{M}_u$ is selected from among $\mathbf{M}_1, \ldots, \mathbf{M}_\ell$ using the associated probabilities. Then for $i = 1, \ldots, M$, $j = 1, \ldots, N$: 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 $m_{121} = 1$ and (1,2,1) is the only unit in D^*_{12}; but, impossibly, $(1,2,1) \in S'_2$ also because $m_{421} = 1$ and (1,2,1) is the only unit in D_{22} for which $\pi'_{i2k2} > 0$.

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 $m_{411} = 0$. 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 these two units would be selected to be in S'_2 since $m_{412} = 1$. The Ernst (1996) procedure selects the sample units corresponding to each cell independently. This does not work here since the same 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 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 each controlled rounding is used together with a probability mechanism to select a sample. The approach used here was chosen to avoid the third problem of Section 2.

The algorithm described in Sections 3.1. and 3.2. results in a single sample. However, what we need is a set of samples, $S'_{1u}, S'_{2u}, S'_{3u}, S'_{4u}$, $u = 1, \ldots, \ell$, and associated probabilities, $p_1, \ldots, p_\ell$, 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, a possible set of samples is given in Table 3 at the end of Section 3.3 for the example in Section 2. Here $\ell = 4$. For each unit (i,j,k) and each sample u, the table entry is that β for which $(i,j,k) \in S'_{\beta u}$. Furthermore, the probabilities associated with these four samples are .4,.2,.2,.2, respectively. For example, (1,2,1) is in sample for both designs for Sample 1, is in sample for neither design for Sample 2, and is in sample for D_2 only for Samples 3 and 4. Thus $\Pr((1,2,1) \in S'_\beta)$ is 0,.4,.4,.2 for $\beta = 1, 2, 3, 4$, respectively, in agreement with (2.4). 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 subsets, namely:

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

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

$$T_3 = \{(i,j,k) : \pi_{ijk1} = \pi_{ijk2}\}$$

In our example, $T_{1C} = \varnothing$, $T_{1S} = \{(3,1,1)\}$, $T_{2C} = \{(1,1,1)\}$, $T_{2S} = \{(1,1,2),(1,2,1)\}$, $T_3 = \{(2,1,1),(2,2,1),(2,2,2),(3,2,1)\}$. We also let $T_\alpha = T_{\alpha C} \cup T_{\alpha S}$, $\alpha = 1, 2$.

As we will show, the partitioning by the numerical subscript overcomes the first problem of Section 2 and the further partitioning determined by the C and S subscripts 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 $\mathbf{A}$ of dimensions $(M+2) \times (N+2)$ described in Section 2. The expanded $\mathbf{A}$ contains five subarrays corresponding to the five sets in the partition of T. 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 the array $\mathbf{A}$ of Section 2, except that each subarray is restricted to the units in the corresponding subset. Furthermore, $\mathbf{A}_{1C}, \mathbf{A}_{1S}$ have dimensions $M \times (N+1)$ instead of dimensions $(M+1) \times (N+1)$; $\mathbf{A}_{2C}$, $\mathbf{A}_{2S}$ 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 allow 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. The expanded array **A** for the example of Section 2 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 column and row numbers, respectively, of **A**. In all subsequent arrays in the article those rows and columns consisting entirely of zeros are omitted to conserve space.

Let $T_{ij1C} = \{k : (i,j,k) \in T_{1C}\}, i = 1,\ldots,M, j = 1,\ldots,N$, with analogous definitions for $T_{ij1S}, T_{iJ2C}, T_{ij2S}, T_{ij3}, T_{ij1}, T_{ij2}$. $\mathbf{A}_3$ occupies the upper left-hand corner of **A** with its elements defined by

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

$\mathbf{A}_{2C}$ 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 $M+2$, in order that these two subarrays shall 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 is to insure that none of the other subarrays has 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 $M+1$ of $\mathbf{A}_{2C}$, the row to be used in selecting units in S'_2, we let

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

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

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

while for Column $N+1$ of $\mathbf{A}_{1S}$, M is replaced by $2M$ on the left-hand side of (3.3) and C is replaced 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 those elements that are in both the first ten rows and first seven columns of the array **A** at the end of this subsection**.**

M is a controlled rounding of **A** and is used to select the first sample. We proceed to 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 define and explain the need for the final N internal rows in $\mathbf{A}$. The definition and explanation for the final M internal columns is analogous. Let

$$a_{j2}' = \sum_{i=1}^{3M+1} a_{ij},\ m_{j2}' = \sum_{i=1}^{3M+1} m_{ij},\ j = 1, \ldots, 3N+1 \tag{3.4}$$

$$a_{j2}'' = a_{j2}' + a_{(j+N+1)2}' + a_{(j+2N+1)2}',\ m_{j2}'' = m_{j2}' + m_{(j+N+1)2}' + m_{(j+2N+1)2}',\ j = 1, \ldots, N \tag{3.5}$$

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 $D_{j2} \cap (T_3 \cup T_1)$, $D_{j2} \cap T_{2C}$, $D_{j2} \cap T_{2S}$, respectively, to be selected to be in S_2, $j = 1, \ldots, N$; 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 $j = 1, \ldots, N$,

$$a_{j2}' = \sum_{i=1}^{M} \sum_{k \in T_{ij3} \cup T_{ij1}} \pi_{ijk2},\ a_{(j+N+1)2}' = \sum_{i=1}^{M} \sum_{k \in T_{ij2C}} \pi_{ijk2},\ a_{(j+2N+1)2}' = \sum_{i=1}^{M} \sum_{k \in T_{ij2S}} \pi_{ijk2} \tag{3.6}$$

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 $m_{j2}'' = a_{j2}''$ by (1.1). To force this last relationship to be true for any controlled rounding $\mathbf{M}$ of $\mathbf{A}$, we define elements in the last N internal rows of $\mathbf{A}$ as follows. For any real number x, let $\lceil x \rceil$ be the smallest integer that is larger than or equal to x and let $r(x) = \lceil x \rceil - x$. Then let

$$a_{(j+3M+1)j} = r(a_{j2}'),\ a_{(j+3M+1)(j+2N+1)} = r(a_{(j+2N+1)2}'),\ j = 1, \ldots, N \tag{3.7}$$

and let the cell value be 0 for all other internal cells in Row $j + 3M + 1$ of $\mathbf{A}$. It is established in Section 6.1 that $m_{j2}'' = a_{j2}''$.

The entries in the final M internal columns of $\mathbf{A}$ are defined analogously to entries in the final N internal rows, that is,

$$a_{i1}' = \sum_{j=1}^{3N+1} a_{ij},\ i = 1, \ldots, 3M+1 \tag{3.8}$$

$$a_{i(i+3N+1)} = r(a_{i1}'),\ a_{(i+2M+1)(i+3N+1)} = r(a_{(i+2M+1)1}'),\ i = 1, \ldots, M \tag{3.9}$$

and the cell value is 0 for all other internal elements in Column $i + 3N + 1$ of $\mathbf{A}$. (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}, $i = 1, \ldots, M$.

This completes the definition of the internal elements of $\mathbf{A}$. The remaining elements are the marginals. $\mathbf{M}$ is any controlled rounding of $\mathbf{A}$. The complete array $\mathbf{A}$ for the example

and a possible **M** are given below.

$\mathbf{A} =$

	1	2	3	4	5	6	7	8	9	10	11
1	0	0	0	.4	0	.2	.4	0	0	0	1
2	.4	.6	0	0	0	0	0	0	0	0	1
3	0	6	0	0	0	0	0	0	0	.4	1
4	0	0	0	.6	0	.2	.4	0	0	0	1.2
5	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	0
10	.2	0	.2	0	0	0	0	0	0	.6	1
11	.4	0	0	0	0	.6	0	0	0	0	1
12	0	.8	0	0	0	0	.2	0	0	0	1
13	1	2	.2	1	0	1	1	0	0	1	7.2

$\mathbf{M} =$

	1	2	4	6	7	10	11
1	0	0	0	0	1	0	1
2	1	0	0	0	0	0	1
3	0	1	0	0	0	0	1
4	0	0	1	0	0	0	1
10	0	0	0	0	0	1	1
11	0	0	0	1	0	0	1
12	0	1	0	0	0	0	1
13	1	2	1	1	1	1	7

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

$$\text{if } \pi'_{ijk3} = 0 \text{ then } (i,j,k) \text{ cannot be in } S_3' \text{ and if } \pi'_{ijk3} = 1 \text{ then } (i,j,k) \text{ must be in } S_3'. \tag{3.10}$$

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 the inequality

$m_{ij} \le \lceil a_{ij} \rceil$, while the second condition follows from (3.1) and the inequality $\lfloor a_{ij} \rfloor \le m_{ij}$. We select the units to be in S'_3 from the corresponding cells of $\mathbf{A}_{1C}, \mathbf{A}_{1S}, \mathbf{A}_{2C}, \mathbf{A}_{2S}$ in the same way. Note, however, that it is not possible for $\pi'_{ijk3} = 1$ if $(i,j,k) \notin T_3$.

After all the units to be in S'_3 are selected, units are selected from T_{2C}, T_{2S} corresponding to the cells in the last row of $\mathbf{A}_{2C}, \mathbf{A}_{2S}$, respectively, 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 $\mathbf{A}$, 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 $\mathbf{A}_{2S}$.

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.

In Section 6.2. we show that for this selection method and $\alpha = 1, 2$ there are a sufficient number of units in T_α not selected to be in S'_3 to select the required number of units to be in S'_α, and, consequently,

$$\text{if } (i,j,k) \in T_\alpha \text{ then } (i,j,k) \notin S'_\alpha \cap S'_3 \tag{3.11}$$

In addition, we show that for any $(i,j,k) \in T$

$$\text{if } \pi_{ijk\alpha} = 1 \text{ then } (i,j,k) \in S_\alpha \tag{3.12}$$

Consequently, by (3.11), (3.12), the selection avoids Problems 1 and 2 of Section 2.

Furthermore, for this selection method, if a unit is not in T_α then clearly it cannot be in S'_α, $\alpha = 1, 2$. It follows from this result and (3.11) that a unit can be in at most one of S'_1, S'_2, S'_3. Finally, by definition, a unit is in S'_4 if and only if it is not in any of S'_1, S'_2, S'_3 and, therefore, each unit is in exactly one of S'_1, S'_2, S'_3, S'_4.

The particular first sample chosen for the example is given in the first row of Table 3.

3.3. *The recursive process of selecting a set of samples*

The selection of the ℓ samples $S'_{1u}, S'_{2u}, S'_{3u}, S'_{4u}$ and associated probabilities p_u, $u = 1, \ldots, \ell$, described earlier, is done recursively as follows. To obtain sample u and p_u we begin with a set of probabilities $\pi'_{ijk\beta u}, (i,j,k) \in T, \beta = 1,2,3,4$, for $u \ge 1$, and a set of probabilities $p_1, \ldots, p_{u-1}$ for $u > 1$. For $u = 1$ we let $\pi'_{ijk\beta 1} = \pi'_{ijk\beta}$ for all i,j,k,β, which will result in the same possible samples for $u = 1$ as the possible samples for the selection described in Section 3.2. For all u the $\pi'_{ijk\beta u}$ must satisfy the conditions satisfied by the $\pi'_{ijk\beta}$, that is:

$$0 \le \pi'_{ijk\beta u} \le 1 \text{ for all } i,j,k,\beta \tag{3.13}$$

$$\sum_{\beta=1}^{4} \pi'_{ijk\beta u} = 1 \text{ for all } i,j,k \tag{3.14}$$

$$\text{For each } i,j,k, \text{ either } \pi'_{ijk1u} = 0 \text{ or } \pi'_{ijk2u} = 0 \tag{3.15}$$

$$\sum_{i=1}^{M} \sum_{k=1}^{t_{ij}} \pi_{ijk2u} = n_{j2}, \quad j = 1, \ldots, N, \text{ and } \sum_{j=1}^{N} \sum_{k=1}^{t_{ij}} \pi_{ijk1u} = n_{i1},\ i = 1, \ldots, M \tag{3.16}$$

where

$$\pi_{ijk\alpha u} = \pi'_{ijk\alpha u} + \pi'_{ijk3u}, \alpha = 1, 2 \tag{3.17}$$

For $u = 1$, (3.13)–(3.16) are satisfied since $\pi'_{ijk\beta 1} = \pi'_{ijk\beta}$. For general u we assume that $\pi'_{ijk\beta u}$ satisfies these relations; proceed to explain how sample u is selected and p_u calculated; define the set of $\pi'_{ijk\beta(u+1)}$ in terms of sample u, the set of $\pi'_{ijk\beta u}$, and $p_1, \ldots, p_u$; establish that the set of $\pi'_{ijk\beta(u+1)}$ satisfy (3.13)–(3.16); and finally explain how the recursive process terminates.

To obtain sample u, first an array $\mathbf{A}_u$ is constructed exactly as $\mathbf{A}$ was constructed in Section 3.1 except $\pi'_{ijk\beta u}$ replaces $\pi'_{ijk\beta}$. In particular, in this construction, T_{1C} and the other four subsets that form a partition of T depends on π_{ijk1u}, π_{ijk2u}, not π_{ijk1}, π_{ijk2}; that is, a unit can be in different subsets for different u. To illustrate, the subscripts of the subset for each unit and sample for our example are listed in Table 2.

Next a controlled rounding $\mathbf{M}_u$ of $\mathbf{A}_u$ is obtained and a sample $S'_{1u}, S'_{2u}, S'_{3u}, S'_{4u}$ selected exactly as the sample 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 $p_1, \ldots, p_{u-1}$, and then recursively compute $\pi'_{ijk\beta(u+1)}$ as follows. For $(i,j,k) \in T$, $\beta = 1, 2, 3, 4$, let:

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

$$p^*_u = \min\{\pi''_{ijk\beta u} : (i,j,k) \in T,\ \beta = 1, 2, 3, 4\} \tag{3.19}$$

$$p_u = p^*_u \text{ if } u = 1,\ p_u = \left(1 - \sum_{\gamma=1}^{u-1} p_\gamma\right) p^*_u \text{ if } u > 1 \tag{3.20}$$

$$\lambda_{ijk\beta u} = 1 \text{ if } (i,j,k) \in S'_{\beta u},\ \lambda_{ijk\beta u} = 0 \text{ if } (i,j,k) \notin S'_{\beta u} \tag{3.21}$$

and, if $p^*_u < 1$

$$\pi'_{ijk\beta(u+1)} = \frac{\pi'_{ijk\beta u} - \lambda_{ijk\beta u} p^*_u}{1 - p^*_u} \tag{3.22}$$

In Section 6.3, we show that if (3.13)–(3.16) are satisfied for u then they are satisfied for $u + 1$.

The recursive process eventually terminates since, as is established in Section 6.4,

$$\text{there is an integer } \ell \text{ for which } p^*_\ell = 1 \tag{3.23}$$

Table 2. Subset of T for each unit by sample

	(i,j,k)							
u	(1,1,1)	(1,1,2)	(1,2,1)	(2,1,1)	(2,2,1)	(2,2,2)	(3,1,1)	(3,2,1)
1 and 2	2*C*	2*S*	2*S*	3	3	3	1*S*	3
3	2*C*	3	2*C*	3	3	3	1*C*	3
4	3	3	2*C*	3	3	3	3	3

Table 3. β for which $(i,j,k) \in S'_{\beta u}$ for each unit and sample

	(i,j,k)							
u	(1,1,1)	(1,1,2)	(1,2,1)	(2,1,1)	(2,2,1)	(2,2,2)	(3,1,1)	(3,2,1)
1	2	4	3	3	4	4	4	3
2	3	2	4	4	4	3	4	3
3	2	3	2	4	3	4	1	4
4	3	4	2	4	3	4	3	4

Consequently,

$$p_\ell = 1 - \sum_{u=1}^{\ell-1} p_u \tag{3.24}$$

which ends the algorithm. It is established in Section 6.5 that this set of ℓ samples satisfies

$$\sum_{\gamma=1}^{\ell} \lambda_{ijk\beta\gamma} p_\gamma = \pi'_{ijk\beta},\ (i,j,k) \in T,\ \beta = 1,2,3,4 \tag{3.25}$$

which is equivalent to (2.4).

The results of using the recursive algorithm for the example are presented below and in Table 3 above. Here $\ell = 4$. Arrays $\mathbf{A}_2, \mathbf{M}_2, \mathbf{A}_3, \mathbf{M}_3$, are presented, along with the samples in Table 3 and the $\pi'_{ijk\beta u}$, $\beta = 1,2,3$, in Table 4. $\mathbf{A}_4, \mathbf{M}_4$, which are identical, have been omitted since, as is established in Section 6.4, there is only one possible Sample 4, which is given by (6.7). That is, $\mathbf{A}_4, \mathbf{M}_4$ is the array with cell value 1 for Cells (1,1), (2,2), (3,1), and (4,5), and 0 for all other internal cells. The π'_{ijk4u} are omitted but can be calculated from (2.1). We also have that the p^*_u are .4,.33,.5,1, respectively, and that the p_u are .4,.2,.2,.2, respectively.

In this example the samples are, with one exception, uniquely determined by the $\mathbf{M}_u$ since there was usually no more than one unit eligible to be chosen for each cell. The exception occurred in Sample 2, where we could have chosen either (2,2,1) or (2,2,2) to

Table 4. $\pi'_{ijk\beta u}$

		(i,j,k)							
u	β	(1,1,1)	(1,1,2)	(1,2,1)	(2,1,1)	(2,2,1)	(2,2,2)	(3,1,1)	(3,2,1)
1	1	0	0	0	0	0	0	.2	0
1	2	.6	.2	.4	0	0	0	0	0
1	3	.4	.2	.4	.4	.4	.2	.2	.6
2	1	0	0	0	0	0	0	.33	0
2	2	.33	.33	.67	0	0	0	0	0
2	3	.67	.33	0	0	.67	.33	.33	.33
3	1	0	0	0	0	0	0	.5	0
3	2	.5	0	1	0	0	0	0	0
3	3	.5	.5	0	0	1	0	.5	0
4	1	0	0	0	0	0	0	0	0
4	2	0	0	1	0	0	0	0	0
4	3	1	0	0	0	1	0	1	0

be in S_3' since $m_{222} = 1$. We chose (2,2,2).

$\mathbf{A}_2 =$

	1	2	3	4	6	7	10	11
1	0	0	0	.67	.33	0	0	1
2	0	1	0	0	0	0	0	1
3	0	.33	0	0	0	0	.67	1
4	0	0	0	.33	.33	.67	0	1.33
10	.33	0	.33	0	0	0	.33	1
11	.67	0	0	0	.33	0	0	1
12	0	.67	0	0	0	.33	0	1
13	1	2	.33	1	1	1	1	7.33

$\mathbf{M}_2 =$

	1	2	4	6	7	10	11
1	0	0	1	0	0	0	1
2	0	1	0	0	0	0	1
3	0	1	0	0	0	0	1
4	0	0	0	1	0	0	1
10	0	0	0	0	0	1	1
11	1	0	0	0	0	0	1
12	0	0	0	0	1	0	1
13	1	2	1	1	1	1	7

$\mathbf{A}_3 =$

	1	2	3	4	5	11
1	.5	0	0	.5	0	1
2	0	1	0	0	0	1
4	0	0	0	.5	1	1.5
7	.5	0	.5	0	0	1
13	1	1	.5	1	1	4.5

$\mathbf{M}_3 =$

	1	2	3	4	5	11
1	1	0	0	0	0	1
2	0	1	0	0	0	1
4	0	0	0	1	1	2
7	0	0	1	0	0	1
13	1	1	1	1	1	5

4. Minimization of Overlap

The procedure described in the previous section can be modified to minimize overlap instead of maximizing overlap by making the following changes. While S_1 remains unchanged, let S_2 denote the random set consisting of the units *not* in sample for D_2 instead of the units in sample for D_2. Analogously, let π_{ijk2} denote the probability that unit (i, j, k) is not in sample for D_2 and n_{j2} denote the number of units in D_{j2} not in sample for D_2. The definitions in (2.1) remain unchanged, as does the method for selecting the samples and associated probabilities described in Section 3.

With these changes, (2.4) still holds, exactly as it does for the maximization problem, and (2.1) and (2.4) imply (2.2). Now, unlike in the previous sections, (2.2) with $\alpha = 2$ states that for this sampling procedure $\Pr(i,j,k) \in S_2$ is the predetermined probability, π_{ijk2}, that this unit is not in sample for D_2, but this immediately implies that $\Pr(i,j,k) \notin S_2$ is the predetermined probability, $1 - \pi_{ijk2}$, that this unit is in sample for D_2.

As for the minimization condition, note that with S_2 defined as in this section, it is now S_1' that is the set of units that are in both samples, rather than S_3' as in the maximization problem. Thus, we want $\Pr((i,j,k) \in S_1')$ to be minimal for each $(i,j,k) \in T$. Now by (2.1) and (2.2),

$$\Pr((i,j,k) \in S_1') = \pi_{ijk1} - \min\{\pi_{ijk1}, \pi_{ijk2}\} = \max\{0, \pi_{ijk1} - \pi_{ijk2}\} \tag{4.1}$$

This is clearly the minimum possible value for (i,j,k) to be in both samples, since if the probability, π_{ijk1}, that (i,j,k) is in sample for D_1 exceeds the probability, π_{ijk2}, that it is not in sample for D_2, then the probability that it is in both samples must be at least the difference of the these two values.

5. A Simulation Study

In the first part of this two-part study, we used the procedure of Section 3 to maximize overlap when selecting two samples of 310 establishments from two different designs for each of 22 industry strata from an artificial universe consisting of 2,736 establishments created for a previous project (Springer et al. 1999) at the U.S. Bureau of Labor Statistics (BLS). The frame for the study was developed to imitate the universe for a middle-sized metropolitan area. The industry strata were identical for the two designs and thus the problem can be viewed as 22 separate overlap problems. The two designs correspond in part to the designs for two BLS compensation surveys. The Ernst (1998) procedure was originally developed to implement a plan to maximize the overlap of sample establishments for a portion of these two compensation surveys, one of which has since been replaced. This application is described in detail in that paper.

For the D_1 design the sampling within each industry stratum was PPS. The universe sizes in the industry strata ranged from eleven to 426 and the sample sizes ranged from one to 46. For the D_2 design we partitioned each industry stratum into seven size class substrata and allocated the industry sample among the substrata proportional to the aggregate measure of size. Thus for each industry stratum we have $M = 1$, $N = 7$, and the dimensions of $\mathbf{A}$ are 12×24.

The results of this simulation are as follows. The expected number of units in sample for both designs over all 22 industry strata when using the overlap procedure is 287.3 out of a total sample of 310, which is a 92.7% overlap. In comparison, if the two samples were selected independently then the expected overlap would be 111.6 or a 36.0% overlap. The CPU time, on a SUN E4500 computer, for solving these 22 overlap problems ranged from two to 249 seconds and the number of possible samples, ℓ, ranged from seven to 807.

A second simulation was run to gauge the computational efficiency of the procedure on a larger problem. In this simulation, the D_1 design was identical to the D_1 design in the first simulation. For the D_2 design, however, this time no size class substrata were used, but instead the sampling within the 22 industry strata was PPS as in the D_1 design. However, in order that the two designs not be identical, the establishments were randomly assigned

to different industry strata in the D_2 design in a way that did not change the number of establishments in the universe for any of the strata. The sample sizes for the D_2 strata were the same as for the D_1 strata. Thus, since the universes for the industry strata are now different for the two designs, we now have a single relatively large overlap problem rather than 22 smaller problems. Here $M = 22$, $N = 22$ and the dimensions of $\mathbf{A}$ are 90×90.

The results of the second simulation are as follows. The expected number of units in sample for both designs over all 22 industry strata when using the overlap procedure is 261.6 out of a total sample of 310, which is an 84.4% overlap, compared to 99.1 units or a 32.0% overlap if the two samples were selected independently. The CPU time for the overlap problem was 5 hours, 16 minutes and the number of samples was 4,435.

6. Appendix

We prove here some of the claims made in Section 3.

6.1. Proof that $m''_{j2} = a''_{j2}$

We first observe that $a_{M^*(j+N+1)} = a'_{(j+N+1)2}$, $j = 1, \ldots, N$, is an integer by (3.6) and the fact that $\pi_{ijk2} = 1$ for all $(i,j,k) \in T_{2C}$. Furthermore, $a_{M^*j}, a_{M^*(j+2N+1)}$ are integers by (3.7). Then, since a''_{j2} is also an integer by (3.5), and

$$a_{(j+3M+1)N^*} = a_{M^*j} + a_{M^*(j+N+1)} + a_{M^*(j+2N+1)} - a''_{j2} \tag{6.1}$$

this row total is an integer. Consequently, since controlled roundings round integers to themselves we have by (3.5), (6.1),

$$m''_{j2} = m_{M^*j} + m_{M^*(j+N+1)} + m_{M^*(j+2N+1)} - m_{(j+3M+1)N^*} = a''_{j2}$$

6.2. Proof of (3.11), (3.12)

We will establish these relations for $\alpha = 2$. The proof for $\alpha = 1$ is similar. We will consider units in T_{2C}, T_{2S} separately for (3.11) and units in T_{2C}, T_{2S}, T_1, T_3 separately for (3.12).

For cells corresponding to units in $D_{j2} \cap T_{2C}$ we have by (3.4) and the fact that $a'_{(j+N+1)2}$ is an integer by (3.6), that

$$m_{(M+1)(j+N+1)} + \sum_{i=1}^{M} m_{i(j+N+1)} = m'_{(j+N+1)2} = a'_{(j+N+1)2},\ j = 1, \ldots, N \tag{6.2}$$

Furthermore, by (3.6), $a'_{(j+N+1)2}$ is the number of units in $D_{j2} \cap T_{2C}$, which together with (6.2) establish that there are exactly $m_{(M+1)(j+N+1)}$ units in $D_{j2} \cap T_{2C}$ that would not be selected to be in S'_3, all of which would be chosen to be in S'_2. Therefore, all units in $D_{j2} \cap T_{2C}$ would be selected to be in S_2 but none would be selected to be in both S'_2 and S'_3. Thus the units in T_{2C} satisfy (3.11), (3.12).

As for units in $D_{j2} \cap T_{2S}$, we have by (3.4), (3.7) that

$$m_{(M+1)(j+2N+1)} + \sum_{i=1}^{M} m_{i(j+2N+1)} = m'_{(j+2N+1)2} \leq m_{M^*(j+2N+1)} = \lceil a'_{(j+2N+1)2} \rceil \tag{6.3}$$

and also, by (3.6), that there are at least $\lceil a'_{(j+2N+1)2} \rceil$ units in $D_{j2} \cap T_{2S}$. Consequently, there are at least $m_{(M+1)(j+2N+1)}$ units in $D_{j2} \cap T_{2S}$ not selected to be in S'_3; any of these units can be chosen to be in S'_2, with the selection thus satisfying (3.11) for units in T_{2S} and, consequently, (3.11) is established. Since $\pi_{ijk2} \neq 1$ for units in T_{2S} or T_1, these units automatically satisfy (3.12). Finally, all units in T_3 satisfy (3.12) by (3.10).

6.3. *Proof that if (3.13)–(3.16) hold for u then they hold for u + 1*

To prove the results we first extend a previous result by showing that (3.10) holds for all $\beta = 1, 2, 3, 4$, not only $\beta = 3$; that is, for all $(i, j, k) \in T$, $\beta = 1, 2, 3, 4$

$$\text{if } \pi'_{ijk\beta} = 0 \text{ then } (i,j,k) \notin S'_\beta \text{ and if } \pi'_{ijk\beta} = 1 \text{ then } (i,j,k) \in S'_\beta \tag{6.4}$$

We first consider the case when $\pi'_{ijk\beta} = 0$. For $\beta = 3$, this part of (6.4) holds by (3.10). For $\beta = 1, 2$, it holds since if $\pi'_{ijk\beta} = 0$ then $(i, j, k) \notin T_\beta$ and hence $(i, j, k) \notin S'_\beta$ by the method of selecting units in S'_β described in Section 3.2. Finally, if $\pi'_{ijk4} = 0$, then for either $\alpha = 1$ or $\alpha = 2$, $\pi_{ijk\alpha} = 1$ by (2.1) and, consequently, $(i, j, k) \in S_\alpha$ by (3.12).

If $\pi'_{ijk\beta} = 1$ then $\pi'_{ijk\gamma} = 0$ for all $\gamma \neq \beta$ by (2.1) and, consequently, $(i, j, k) \notin S'_\gamma$. Therefore $(i, j, k) \in S'_\beta$ and the proof of (6.4) is complete.

Also observe that if (3.13)–(3.16) holds for u, then Sample u has the same properties as Sample 1, and hence the analog of (6.4) holds for Sample u; that is,

$$\text{if } \pi'_{ijk\beta u} = 0 \text{ then } (i,j,k) \notin S'_{\beta u} \text{ and if } \pi'_{ijk\beta u} = 1 \text{ then } (i,j,k) \in S'_{\beta u} \tag{6.5}$$

We will use this result both in this and the next subsection.

Now we establish the main results of this subsection, that is, if (3.13)–(3.16) hold for u then these relations hold for $u + 1$. We establish this by combining the inductive assumption that they hold for u with (3.22) and the following additional relations.

For (3.13): $p^*_u \leq \pi'_{ijk\beta u}$ if $\lambda_{ijk\beta u} = 1$, and $\pi'_{ijk\beta u} \leq 1 - p^*_u$ if $\lambda_{ijk\beta u} = 0$, which follow from (3.18), (3.19).

For (3.14): the relation $\Sigma^4_{\beta=1} \lambda_{ijk\beta u} = 1$, which states that each unit is in exactly one of $S'_{1u}, S'_{2u}, S'_{3u}, S'_{4u}$.

For (3.15): $\lambda_{ijk\alpha u} = 0$ if $\pi'_{ijk\alpha u} = 0$, which follows from (6.5). Note that the same reasoning can also be used to establish the more general result that

$$\text{if } \pi'_{ijk\beta u} = 0 \text{ or } \pi'_{ijk\beta u} = 1 \text{ for some } i, j, k, \beta, u, \text{ then } \pi'_{ijk\beta(u+1)} = \pi'_{ijk\beta u} \tag{6.6}$$

which we will use in the next subsection.

For the first equation in (3.16): (3.17) and the relation $\Sigma^M_{i=1} \Sigma^{t_{ij}}_{k=1} (\lambda_{ijk2u} + \lambda_{ijk3u}) = n_{j2}$, $j = 1, \ldots, N$, which follows from $m''_{ju2} = n_{j2}$, where m''_{ju2} is the analog of m''_{j2} for Sample u. The second equation in (3.16) is established similarly.

6.4. *Proof of (3.23)*

By (3.18), (3.19), and (6.5), for each u we have $p^*_u \leq 1$ and also $p^*_u < 1$ if and only if there is at least one i, j, k, β for which $0 < \pi'_{ijk\beta u} < 1$. Furthermore, if $p^*_u < 1$ then there is some i, j, k, β, for which $0 < \pi'_{ijk\beta u} < 1$ and $\pi''_{ijk\beta u} = p^*_u$. By (3.18), (3.21), and (3.22), $\pi'_{ijk\beta(u+1)} = 0$ or $\pi'_{ijk\beta(u+1)} = 1$ for this i, j, k, β, which together with (6.6) establish that

$\{i,j,k,\beta : \pi'_{ijk\beta u} = 0 \text{ or } \pi'_{ijk\beta u} = 1\}$ is a strictly increasing set as a function of u. Consequently, $\pi'_{ijk\beta\ell} = 0$ or $\pi'_{ijk\beta\ell} = 1$ for all i,j,k,β for some ℓ. Furthermore, by (3.14), $\pi'_{ijk\beta\ell} = 1$ for each i,j,k, for exactly one β. Then by (6.5) there is only one possible sample ℓ, namely the sample for which

$$\lambda_{ijk\beta\ell} = \pi'_{ijk\beta\ell} \text{ for all } i,j,k,\beta \tag{6.7}$$

Finally, $p^*_\ell = 1$ by (6.7), (3.18), (3.19).

6.5. *Proof of (3.25)*

We establish by induction that

$$\pi'_{ijk\beta(u+1)}\left(1 - \sum_{\gamma=1}^{u} p_\gamma\right) + \sum_{\gamma=1}^{u} \lambda_{ijk\beta\gamma} p_\gamma = \pi'_{ijk\beta},\ u = 1,\ldots,\ell-1 \tag{6.8}$$

and then combine (6.8) for $u = \ell - 1$ with (6.7), (3.24) to conclude (3.25). For $u = 1$, (6.8) follows immediately from (3.22) with the substitutions $p^*_1 = p_1$, $\pi'_{ijk\beta 1} = \pi'_{ijk\beta}$. If (6.8) holds with u replaced by $u-1$ then it holds for u since if we solve (3.22) for $\pi'_{ijk\beta u}$, substitute the result in (6.8) with u replaced by $u-1$, and use (3.20), we obtain

$$\begin{aligned}\pi'_{ijk\beta} &= \pi'_{ijk\beta u}\left(1 - \sum_{\gamma=1}^{u-1} p_\gamma\right) + \sum_{\gamma=1}^{u-1} \lambda_{ijk\beta\gamma} p_\gamma = \pi'_{ijk\beta(u+1)}\left(1 - \sum_{\gamma=1}^{u-1} p_\gamma\right) \\ &\quad + (\lambda_{ijk\beta u} - \pi'_{ijk\beta(u+1)})\left(1 - \sum_{\gamma=1}^{u-1} p_\gamma\right) p^*_u + \sum_{\gamma=1}^{u-1} \lambda_{ijk\beta\gamma} p_\gamma \\ &= \pi'_{ijk\beta(u+1)}\left(1 - \sum_{\gamma=1}^{u} p_\gamma\right) + \sum_{\gamma=1}^{u} \lambda_{ijk\beta\gamma} p_\gamma\end{aligned}$$

7. References

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.

Cox, L. H. and Ernst, L. R. (1982). Controlled Rounding. INFOR, 20, 423–432.

Ernst, L. R. (1996). Maximizing the Overlap of Sample Units for Two Designs with Simultaneous Selection. Journal of Official Statistics, 12, 33–45.

Ernst, L. R. (1998). Maximizing and Minimizing Overlap When Selecting a Large Number of Units per Stratum Simultaneously for Two Designs. Journal of Official Statistics, 14, 297–314.

Ernst, L. R. (1999). The Maximization and Minimization of Sample Overlap Problems: A Half Century of Results. International Statistical Institute, Proceedings, Invited Papers, IASS Topics, 168–182.

Goodman, R. and Kish, L. (1950). Controlled Selection—A Technique in Probability Sampling. Journal of the American Statistical Association, 45, 350–372.

Keyfitz, N. (1951). Sampling With Probabilities Proportionate to Size: Adjustment for Changes in Probabilities. Journal of the American Statistical Association, 46, 105–109.

Springer, G., Walker, M., Paben, S., and Dorfman, A. (1999). Evaluation of Confidence Interval Methodology for the National Compensation Survey. Proceedings of the American Statistical Association, Section on Survey Research Methods, 486–491.

Received May 2001
Revised March 2002