Journal of Official Statistics, Vol. 12, No. 1, 1996, pp. 33–45

Maximizing the Overlap of Sample Units for Two Designs with Simultaneous Selection

Lawrence R. Ernst[1]

It is demonstrated, using transportation theory, that controlled selection can be used to solve the following sampling problem. Sample units are to be selected with probability proportional to size for two designs, both one unit per stratum, denoted as D_1 and D_2, with generally different stratifications. The goal of the problem is to simultaneously select the sample units for the two designs in a manner which maximizes the expected number of units that are in both samples. The procedure differs from previous overlap procedures in that it yields a better overlap, but is only applicable when the two samples can be selected simultaneously. An important special case occurs when the probability of selection for each unit in D_1 does not exceed its probability of selection in D_2. The procedure can then guarantee that the D_1 sample units are a subset of the D_2 sample units. A proposed, but since canceled, expansion of the Current Population Survey, which is discussed, would have been a potential application of this special case. Variance formulas for estimators of total under the controlled selection procedure are also presented. In addition, it is demonstrated that the procedure can easily be modified to minimize expected overlap instead of maximizing it.

Key words: Controlled selection; Current Population Survey; overlap maximization; stratification.

1. Introduction

Consider the following sampling problem: Sample units are to be selected for two designs, denoted as D_1 and D_2, both of which are one unit per stratum designs. (Typically, the units are actually primary sampling units (PSUs) in a multistage design.) The selection of sample units for each design is to be with probability proportional to a measure of size which need not be the same for the two designs. The universes of sampling units for the two designs have some, but not necessarily all, units in common. The two designs are stratified independently, with the sample units for the two designs then to be selected simultaneously. We wish to maximize the

[1] Bureau of Labor Statistics, Office of Compensation and Working Conditions, Research Group, Washington, D.C. 20212, U.S.A.

Author's Footnote: Currently, Chief, Research Group, Office of Compensation and Working Conditions, Bureau of Labor Statistics, Washington D.C. 20212. Formerly, Assistant Chief, Statistical Research Division, Bureau of the Census, Washington, D.C. 20233.

Acknowledgment: The views expressed in this article are attributable to the author and do not necessarily reflect those of the Bureau of Labor Statistics or the Census Bureau. The programming assistance of Todd Williams is gratefully acknowledged. The author also thanks the referees and the Associate Editor, Bengt Rosén, for their very constructive comments.

overlap of the sample units, that is to select the sample units so that

One unit is selected from each D_1 and each D_2 stratum. (1)
Each unit is selected into each design with the required probability. (2)
The expected value for the number of sample units common to the two designs is maximized. (3)

In this article we demonstrate how the two-dimensional controlled selection procedure of Causey, Cox, and Ernst (1985) can be used to satisfy these conditions and the additional condition that

The number of sample units in common to any D_1 and D_2 samples is always within one of the maximum expected value. (4)

Most of the previous work on maximizing the overlap of sample units considered the case when the two sets of sample units are chosen sequentially. This problem was first studied by Keyfitz (1951), who presented an optimum procedure for one unit per stratum designs in the special case when the initial and new strata are identical, with only the selection probabilities changing. For the more general one unit per stratum problem, Perkins (1970), and Kish and Scott (1971) presented procedures that are not optimal in the sense of (3).

Causey, Cox, and Ernst (1985), and Ernst (1986) presented optimal linear programming procedures for maximizing the expected number of sample units in common to the two designs, under very general conditions, when the two sets of sample units are chosen sequentially. These last two papers impose no restrictions on changes in strata definitions or number of units per stratum. Brewer, Early, and Joyce (1972) considered a somewhat similar problem except, unlike the other authors, they did not fix the sample size, which allows for a much simpler solution.

A typical application of overlap maximization in the sequential case occurs when the two designs are for the same periodic household survey, but the second design is a redesign of the first design done at a later date. The sampling units are PSUs, and the motivation for using an overlap procedure is to reduce additional costs, such as the training of a new interviewer, incurred with each change of a sample PSU. In general, as will be demonstrated in Section 5 of this article, choosing the two samples simultaneously permits a larger expected overlap, but in applications such as the one just described it is not possible to select the samples simultaneously.

Pruhs (1989) was the first to consider the problem of maximizing overlap for simultaneous selection under the conditions to be considered in this article. Using a graph theory approach, he presented an algorithm which satisfies conditions (1)–(4). It is shown here that this problem can also be solved by the controlled selection procedure of Causey, Cox, and Ernst (1985). This approach has two advantages over Pruh's approach. The controlled selection approach involves solving a sequence of transportation problems. Software is readily available to solve a transportation problem, which is a special form of linear programming problem for which extremely efficient solution strategies exist (Glover, Karney, Klingman, and Napier 1974), and the remainder of the controlled selection algorithm is easily programmable. In addition, the proof that the controlled selection procedure satisfies the required conditions is

not difficult. In contrast, both the theory and the task of programming the algorithm with Pruh's graph theory approach is much more complex.

A special case of (1)–(4) occurs when

The universe of sampling units is the same for the two designs and the probability of selection for each unit in D_1 does not exceed its D_2 selection probability. (5)

For this special case it can be shown that (3) and (4) can be replaced with the more stringent requirement that

Each D_1 sample unit is a D_2 sample unit. (6)

That is, all the D_1 sample units overlap with D_2 sample units.

A particular application of the special case to a proposed expansion of the U.S. Current Population Survey (CPS), which was the original motivation for this work, is presented in Section 8. Plans for this expansion have since been dropped for budgetary reasons. Some readers may wish to read the beginning of Section 8 before proceeding further, to obtain an understanding of this motivation.

Most of the overlap procedure is presented in Section 2. The presentation is completed in Section 3, where it is shown how the controlled selection algorithm of Causey, Cox, and Ernst (1985) can be used to obtain a key step in the procedure. In Section 4, variance formulas for this procedure are obtained for both designs for the usual estimator of total corresponding to probability proportional to size sampling. In Section 5, as noted previously, it is explained why a higher maximal expected overlap can generally be obtained by simultaneous selection of the sample units for the two designs than by sequential selection.

In Section 6 it is shown how the procedure of Sections 2 and 3 can be easily modified to solve the problem of minimizing the expected overlap of sample units under the same assumptions. Perry, Burt, and Iwig (1993) have recently presented a different approach to the minimization of overlap when the samples are selected simultaneously. Their approach has the advantage of not being restricted to two designs or one unit per stratum. However, their method is not optimal and assumes equal probability of selection within a stratum.

In Section 7 it is explained why the main result of this article is not readily generalizable to other than one unit per stratum designs. Finally, in Section 8, an application of this procedure to the proposed expansion of the CPS is briefly considered.

2. The Maximization of Overlap Algorithm

The sample selection process for this procedure is a two step process. The first step is the selection, by a probability mechanism to be described, of an $(m+2) \times (n+2)$ array, $\tilde{n} = (n_{ij})$, where m and n are the number of D_1 and D_2 strata, respectively. The selected array determines for which i, j a unit is selected from the intersection of the ith D_1 stratum and the jth D_2 stratum to be in sample for both designs, and for which D_1 and D_2 strata, units are selected to be in sample for only one of these designs. The selected array always satisfies (1) and (4), and the probability mechanism for selecting the array guarantees that (3) is satisfied.

At the second step the actual sample units are selected subject to the constraints imposed by the selected array, in a manner that will be shown to satisfy (2).

We now outline the remainder of the section. The random array $\tilde{n}$ is first described in greater detail. We then discuss additional constraints on $\tilde{n}$ and its selection process that are required to satisfy (1), (3) and (4), but postpone to Section 3 discussion of how the controlled selection algorithm can be applied to ensure that these constraints are satisfied. Finally, we detail the second step of the procedure and show that, assuming the constraints on $\tilde{n}$ and its selection process are satisfied, the two-step process satisfies (2). At the end of each portion of the presentation we present the appropriate part of the same example as an illustration.

We proceed to describe the random array in greater detail. For $i = 1, \ldots m$, $j = 1, \ldots n$, n_{ij} is a 0,1 variable which indicates the number of units that are to be selected from the intersection of the ith D_1 stratum and the jth D_2 stratum to be in sample for both designs. Since the two designs are in general different, some units may have to be in sample for only one design. Here, $n_{i(n+1)}$, $i = 1, \ldots m$, is a 0,1 variable that indicates the number of units in the ith D_1 stratum that are to be D_1 only sample units, and similarly $n_{(m+1)\,j}$, $j = 1, \ldots, n$, is the number of units in the jth D_2 stratum that are to be D_2 only sample units. For completeness of the matrix we set $n_{(m+1)(n+1)} = 0$. The remaining entries in the array are marginals, that is

$$n_{i(n+2)} = \sum_{j=1}^{n+1} n_{ij}, \qquad i = 1, \ldots, m+2, \tag{7}$$

$$n_{(m+2)j} = \sum_{i=1}^{m+1} n_{ij}, \qquad j = 1, \ldots, n+2. \tag{8}$$

That is the array can be represented in the form

n_{11}	. . .	$n_{1(n+1)}$		$n_{1(n+2)}$
.	. . .	.		.
.	. . .	.		.
.	. . .	.		.
$n_{(m+1)1}$	. . .	$n_{(m+1)(n+1)}$		$n_{(m+1)(n+2)}$
$n_{(m+2)1}$	. . .	$n_{(m+2)(n+1)}$		$n_{(m+2)(n+2)}$

with the internal, row total, column total and grand total cells clear from the diagram. An array satisfying the additivity constraints represented by the above diagram is referred to as a tabular array.

We now begin development of the illustrative example. In this example $m = 2$ and $n = 3$. In one solution to the first step for this example there are the following four values for $\tilde{n}$, denoted $\tilde{n}_1 - \tilde{n}_4$

$$\tilde{n}_1 = \begin{array}{cccc|c} 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 \\ \hline 1 & 1 & 1 & 0 & 3 \end{array} \qquad \tilde{n}_2 = \begin{array}{cccc|c} 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 \\ \hline 1 & 1 & 1 & 0 & 3 \end{array}$$

$$\tilde{n}_3 = \begin{array}{cccc|c} 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 \\ \hline 1 & 1 & 1 & 0 & 3 \end{array} \qquad \tilde{n}_4 = \begin{array}{cccc|c} 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 2 \\ \hline 1 & 1 & 1 & 1 & 4 \end{array} \tag{9}$$

The probability that $\tilde{n}_k$ is selected, $k = 1, 2, 3, 4$, is denoted p_k, with $0.5, 0.3, 0.1, 0.1$, respectively, the ordered values of these probabilities for this example. The $\tilde{n}_k$ and their associated probabilities, p_k, constitute a solution to the first step of the procedure. (Generally, a solution to this step is not unique.) Although not enough information has yet been provided to obtain or verify this solution, (9) can be used to illustrate the role of the selected $\tilde{n}$ in the sampling process. For example, if $\tilde{n}_1$ is selected then, since $n_{11} = 1$ and $n_{23} = 1$, one unit is selected from the intersection of D_1 stratum 1 and D_2 stratum 1, and one unit is selected from the intersection of D_1 stratum 2 and D_2 stratum 3, to be in sample for both designs. Also, since $n_{32} = 1$, one unit is selected from D_2 stratum 2 to be in sample only for the D_2 design.

We return to the development of the first step of the procedure. Additional constraints on the array $\tilde{n}$ and its selection mechanism are required in order to satisfy (1), (3) and (4). To satisfy (1) we must have

$$n_{i(n+2)} = 1, \qquad i = 1, \ldots m, \tag{10}$$

and

$$n_{(m+2)j} = 1, \qquad j = 1, \ldots, n. \tag{11}$$

Note that each of the arrays in (9) satisfies these conditions.

Before considering the constraints needed to satisfy (3) and (4) we present some additional notation. For this notation and throughout the remainder of Sections 2 and 3 we need to be able to treat the universes of sampling units for D_1 and D_2 as identical. Since we purposely did not make this assumption in the Introduction, 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.

Let S_1 and S_2 denote the random sets consisting of all sample units in D_1 and D_2, respectively. For $i = 1, \ldots, m$, $j = 1, \ldots, n$, let t_{ij} denote the number of units in the population that are in the intersection of the ith D_1 stratum and the jth D_2 stratum; let B_{ijk} denote the kth such unit, $k = 1, \ldots, t_{ij}$; and let T denote the set of all triples (i,j,k) within the indicated domains for i, j, k. For $(i,j,k) \in T$, let π_{ijk1}, π_{ijk2} denote the preassigned selection probabilities for B_{ijk} in the D_1 and D_2 designs, respectively, and let $\pi_{ijk3} = \min\{\pi_{ijk1}, \pi_{ijk2}\}$. Finally, let

$$s_{ij} = \sum_{k=1}^{t_{ij}} \pi_{ijk3} \qquad i = 1, \ldots, m, \qquad j = 1, \ldots n. \tag{12}$$

Note that (2) is then equivalent to the requirement that

$$P(B_{ijk} \in S_\alpha) = \pi_{ijk\alpha}, \qquad (i,j,k) \in T, \qquad \alpha = 1, 2. \tag{13}$$

For any designs fulfilling (13) we have $P(B_{ijk} \in S_1 \cap S_2) \leqslant \pi_{ijk3}$ for all $(i,j,k) \in T$, which, by (12), implies that $P(n_{ij} = 1) \leqslant s_{ij}$, $i = 1,\ldots,m, j = 1,\ldots n$. Hence (3) will be satisfied if we impose the more restrictive requirement on the random array $\tilde{n}$ that

$$P(n_{ij} = 1) = s_{ij}, \qquad i = 1,\ldots,m, \qquad j = 1,\ldots,n. \tag{14}$$

It would follow from (14) that the expected number of sample units in common to the two designs is

$$E\left(\sum_{i=1}^{m}\sum_{j=1}^{n} n_{ij}\right) = \sum_{i=1}^{m}\sum_{j=1}^{n} s_{ij}. \tag{15}$$

Consequently, in order to establish (4), we impose the further requirement that

$$\left|\sum_{i=1}^{m}\sum_{j=1}^{n} n_{ij} - \sum_{i=1}^{m}\sum_{j=1}^{n} s_{ij}\right| < 1 \tag{16}$$

for each possible value for $\tilde{n}$. In Section 3 we demonstrate how to obtain a set of non-negative integer valued arrays and associated selection probabilities satisfying (7), (8), (10), (11), and (14–16). Below, we demonstrate that these relations together with the method of selecting the units in the second step of the procedure yield (13). All of these relations together immediately imply (1–4).

Also observe that in the special case when (5) holds, then $\pi_{ijk3} = \pi_{ijk1}$ for all $(i,j,k) \in T$. Consequently, by (12) the right hand side of (15) reduces to m. However, if the expected number of units in common to the two designs is m then (6) holds, and thus this special case follows from the general case.

Before proceeding to the development of the second step in the procedure, we continue with our example to illustrate (12) and (14–16). We must first specify values for the t_{ij}, π_{iju1}, and π_{iju2}. In this example, D_1 and D_2 consist of the identical six units with the following t_{ij}: $t_{11} = 2$, $t_{12} = 1$, $t_{13} = 0$, $t_{21} = 0$, $t_{22} = 1$, $t_{23} = 2$. The selection probabilities for each of these six units for each design are given in Table 1.

It can be computed from this table and (12) , that the 2×3 array (s_{ij}) of desired probabilities that $n_{ij} = 1$ is

$$(s_{ij}) = \begin{pmatrix} 0.8 & 0.1 & 0 \\ 0 & 0.4 & 0.6 \end{pmatrix} \tag{17}$$

and that the expected number of units in common to the two designs must be 1.9 units in order to satisfy (15); consequently, there must always be either 1 unit or 2 units in common to satisfy (16). It can be calculated that the set of 4 arrays in (9) together with their associated probabilities do satisfy (14–16).

Table 1. Selection probabilities for units in example

	ijk					
	111	112	121	221	231	232
π_{ijk1}	0.5	0.4	0.1	0.4	0.4	0.2
π_{ijk2}	0.4	0.6	0.3	0.7	0.6	0.4

We now turn to the second step of the procedure, that is the selection of the sample units conditioned on the chosen array $\tilde{n} = (n_{ij})$. For each i, j, with $i \leqslant m, j \leqslant n$, for which $n_{ij} = 1$, we must have $B_{ijk} \in S_1 \cap S_2$ for a single $k = 1, \ldots, t_{ij}$. The assigned conditional selection probabilities for these t_{ij} units are

$$P(B_{ijk} \in S_1 \cap S_2 | n_{ij} = 1) = \pi_{ijk3}/s_{ij}, \qquad k = 1, \ldots, t_{ij}. \tag{18}$$

In order to assign the conditional selection probabilities for units to be in sample for D_1 only and D_2 only, we first expand the $m \times n$ array (s_{ij}) to an $(m+1) \times (n+1)$ array by letting

$$s_{i(n+1)} = 1 - \sum_{j=1}^{n} s_{ij}, \qquad i = 1, \ldots, m, \tag{19}$$

$$s_{(m+1)j} = 1 - \sum_{i=1}^{m} s_{ij}, \qquad j = 1, \ldots, n, \tag{20}$$

$$s_{(m+1)(n+1)} = 0. \tag{21}$$

Then, for use in the next section, we further expand this array to an $(m+2) \times (n+2)$ tabular array, by adding the marginals

$$s_{i(n+2)} = \sum_{j=1}^{n+1} s_{ij}, \qquad i = 1, \ldots, m+2, \tag{22}$$

$$s_{(m+2)j} = \sum_{i=1}^{m+1} s_{ij}, \qquad j = 1, \ldots, n+2. \tag{23}$$

For example, the 2×3 array (17) expands to the 4×5 tabular array

0.8	0.1	0	0.1	1
0	0.4	0.6	0	1
0.2	0.5	0.4	0	1.1
1	1	1	0.1	3.1

(24)

Now (7), (10), (14) and (19) imply that the probability that a unit is selected from the ith D_1 stratum to be in sample for D_1 only, is

$$P(n_{i(n+1)} = 1) = s_{i(n+1)}, \qquad i = 1, \ldots, m. \tag{25}$$

If $n_{i(n+1)} = 1$ then $B_{ijk} \in (S_1 \sim S_2)$ for a single unit among the $\Sigma_{j=1}^{n} t_{ij}$ units in the ith D_1 stratum. We assign the following conditional selection probabilities

$$P(B_{ijk} \in (S_1 \sim S_2) | n_{i(n+1)} = 1) = (\pi_{ijk1} - \pi_{ijk3})/s_{i(n+1)},$$
$$j = 1, \ldots, n, \qquad k = 1, \ldots, t_{ij}. \tag{26}$$

Similarly, the assigned conditional probabilities for selecting a unit to be in sample only for the jth D_2 stratum when $n_{(m+1)j} = 1$ are

$$P(B_{ijk} \in (S_2 \sim S_1) | n_{(m+1)j} = 1) = (\pi_{ijk2} - \pi_{ijk3})/s_{(m+1)j},$$
$$i = 1, \ldots, m, \qquad k = 1, \ldots, t_{ij}. \tag{27}$$

The conditional selection probabilities just defined yield (13), since for $\alpha = 1$ this follows from (14), (18), (25) and (26) by combining

$$P(B_{ijk} \in S_1 \cap S_2) = P(n_{ij} = 1)P(B_{ijk} \in S_1 \cap S_2 | n_{ij} = 1) = \pi_{ijk3},$$
$$P(B_{ijk} \in (S_1 \sim S_2)) = P(n_{i(n+1)} = 1)P(B_{ijk} \in S_1 \cap S_2 | n_{i(n+1)} = 1) = \pi_{ijk1} - \pi_{ijk3},$$

while (13) for $\alpha = 2$ is obtained similarly.

To illustrate the second step for the example that we have been considering, suppose that the array $\tilde{n}_1$ in (9) is selected at the first step. Then since $n_{11} = 1$, we have by (18) that the conditional probabilities that B_{111}, B_{112} are in sample for both designs are each 1/2. Similarly, since $n_{23} = 1$, the conditional probabilities that B_{231}, B_{232} are in sample for both designs are 2/3 and 1/3, respectively. Finally, since $n_{32} = 1$, the conditional probabilities that B_{121}, B_{221} are in sample for D_2 only are 2/5 and 3/5, respectively, by (27).

3. Controlled Selection

We demonstrate here how the controlled selection procedure of Causey, Cox, and Ernst (1985) can be used to complete the algorithm of this article, that is to construct a finite set of $(m+2) \times (n+2)$ nonnegative, integer-valued, tabular arrays, $\tilde{n}$, and associated probabilities, satisfying (7), (8), (10), (11) and (14–16).

The discussion of controlled selection will be limited to the two-dimensional problem. Although the concept can be generalized to higher dimensions, Causey, Cox, and Ernst (1985) proved that solutions to controlled selection problems do not always exist for dimensions greater than two.

The controlled selection procedure of Causey, Cox, and Ernst is built upon the theory of controlled rounding developed by Cox and Ernst (1982). A controlled rounding of an $(m+2) \times (n+2)$ tabular array (a_{ij}) to a positive integer base b is an $(m+2) \times (n+2)$ tabular array (r_{ij}) for which $r_{ij} = \lfloor a_{ij}/b \rfloor b$ or $(\lfloor a_{ij}/b \rfloor + 1)b$ for all i, j, where $\lfloor x \rfloor$ denotes the greatest integer not exceeding x. A zero-restricted controlled rounding to a base b is a controlled rounding that satisfies the additional condition that $r_{ij} = a_{ij}$ whenever a_{ij} is an integral multiple of b. If no base is specified, then base 1 is understood. As an example, each of the arrays in (9) is a zero-restricted controlled rounding of (24).

By modeling the controlled rounding problem as a transportation problem, Cox and Ernst (1982) obtained a constructive proof that a zero-restricted controlled rounding exists for every two-dimensional array. Thus, while conventional rounding of a tabular array commonly results in an array that is no longer additive, this result shows that it is possible to always preserve additivity if the original values are allowed to be rounded either up or down.

With (a_{ij}) as above, a solution to the controlled selection problem for this array is a finite sequence of $(m+2) \times (n+2)$ tabular arrays, $\tilde{n}_1 = (n_{ij1})$, $\tilde{n}_2 = (n_{ij2})$, ..., $\tilde{n}_l = (n_{ijl})$, and associated probabilities, $p_1, \ldots, p_l$, satisfying

$\tilde{n}_u$ is a zero-restricted controlled rounding of (a_{ij}) for all $u = 1, \ldots, l$, (28)

$$\sum_{u=1}^{l} p_u = 1, \tag{29}$$

$$\sum_{u=1}^{l} n_{iju} p_u = a_{ij}, \qquad i = 1, \ldots, m+2, \qquad j = 1, \ldots, n+2. \tag{30}$$

(Note that in a slight change of notation from Section 2, we use n_{iju} in place of n_{ij} with the additional subscript indicating the uth array, $\tilde{n}_u$.) If (a_{ij}) arises from a sampling problem for which a_{ij} is the expected number of sample units selected in cell (i,j), and the actual number selected in each cell is determined by choosing one of the $\tilde{n}_u$ with its associated probability, then by (28) the deviation of a_{ij} from the number of sample units actually selected from cell (i,j) is less than 1 in absolute value, whether (i,j) is an internal cell or a total cell. By (30) the expected number of sample units selected is a_{ij}.

To illustrate controlled selection, consider the example presented in Section 2. The controlled selection problem for this example is (24). A solution to this problem is the set of arrays presented in (9), together with their associated probabilities.

The concept of controlled selection was first developed by Goodman and Kish (1950), but they did not present a general algorithm for solving such problems. In Causey, Cox, and Ernst (1985), a solution to the controlled selection problem, which will not be reproduced here, was obtained by means of recursive computation of the sequences $\tilde{n}_1, \ldots, \tilde{n}_l$ and $p_1, \ldots, p_l$. We proceed to show that with $(a_{ij}) = (s_{ij})$, a solution to the controlled selection problem satisfies (7), (8), (10), (11) and (14–16).

We see that (7) and (8) follow immediately from (28). Next observe that it follows from (19), (20), (22) and (23) that

$$s_{i(n+2)} = 1, \qquad i = 1, \ldots, m, \quad \text{and} \quad s_{(m+2)j} = 1, \qquad j = 1, \ldots, n \tag{31}$$

which together with (28) yield (10) and (11). To obtain (14), note that for $i = 1, \ldots, m$, $j = 1, \ldots, n$, we have $0 \leqslant s_{ij} \leqslant 1$ by (12), and hence n_{iju} is a 0,1 variable for all u by (28). Then (14) follows from (30) and (15) follows immediately from (14).

To deduce (16) we first obtain from (22), (21) and (20) that

$$s_{(m+1)(n+2)} = \sum_{j=1}^{n} s_{(m+1)j} = n - \sum_{i=1}^{m} \sum_{j=1}^{n} s_{ij}. \tag{32}$$

We next note that $n_{(m+1)(n+1)} = 0$ by (21) and (28), and then combine this result with (7), (8) and (11) to obtain

$$n_{(m+1)(n+2)} = \sum_{j=1}^{n} n_{(m+1)j} = \sum_{j=1}^{n} \left(n_{(m+2)j} - \sum_{i=1}^{m} n_{ij} \right) = n - \sum_{i=1}^{m} \sum_{j=1}^{n} n_{ij}. \tag{33}$$

Finally, we combine (28), (32) and (33) to conclude

$$1 > |n_{(m+1)(n+2)} - s_{(m+1)(n+2)}| = \left| \sum_{i=1}^{m} \sum_{j=1}^{n} (n_{ij} - s_{ij}) \right|.$$

In implementing the controlled selection portion of the selection procedure for the CPS application described in Section 8, some programming difficulties relating to rounding error arose which caused integer-valued marginals, such as (31), to deviate slightly from integer values. These difficulties and the approach used to successfully overcome them are described in detail in Ernst (1993).

4. Variances for the Controlled Selection Procedure

In this section a variance formula is derived for the standard estimator of total for probability proportional to size sampling for D_1 for the sampling procedure detailed in the previous two sections, assuming single stage sampling. The analogous formula for D_2 can be immediately obtained by symmetry. If the units selected by this procedure are actually PSUs for a multistage design, then these formulas are the between PSUs component of variance, in which case formulas for overall variance can be obtained by combining the formula presented here with Raj (1968, p.118).

To compute the variance we simply use the formula in Raj (1968, p. 54, (3.36)), where the summation in the formula is over all distinct pairs of units in D_1, not simply the pairs within the same stratum. The only term in this formula that is not easily computable is the joint probability $P(B_{ijk}, B_{i'j'k'} \in S_1)$, which we denote by $\pi_{ijki'j'k'1}$, for any distinct pair of units in D_1. To compute $\pi_{ijki'j'k'1}$, we first let $r_{iji'j'} = P(n_{iju} = n_{i'j'u} = 1)$ for all i, j, i', j' with $i \leqslant m+1$, $i' \leqslant m+1$, $j \leqslant n+1$, $j' \leqslant n+1$; that is $r_{iji'j'}$ is the sum of p_u over all u for which $n_{iju} = n_{i'j'u} = 1$. Then note that $\pi_{ijki'j'k'1} = 0$ if $i = i'$; while if $i \neq i'$, observe that both B_{ijk} and $B_{i'j'k'}$ can be in S_1 if for some u, either

$$n_{iju} = n_{i'j'u} = 1, \quad n_{i(n+1)u} = n_{i'j'u} = 1, \quad n_{iju} = n_{i'(n+1)u} = 1$$

$$\text{or} \quad n_{i(n+1)u} = n_{i'(n+1)u} = 1$$

which combined with (18) and (26) yield the four terms in the following expression

$$\begin{aligned}\pi_{ijki'j'k'1} = {} & r_{iji'j'} \frac{\pi_{ijk3}}{s_{ij}} \frac{\pi_{i'j'k'3}}{s_{i'j'}} + r_{i(n+1)i'j'} \frac{(\pi_{ijk1} - \pi_{ijk3})}{s_{i(n+1)}} \frac{\pi_{i'j'k'3}}{s_{i'j'}} \\ & + r_{iji'(n+1)} \frac{\pi_{ijk3}}{s_{ij}} \frac{(\pi_{i'j'k'1} - \pi_{i'j'k'3})}{s_{i'(n+1)}} \\ & + r_{i(n+1)i'(n+1)} \frac{(\pi_{ijk1} - \pi_{ijk3})}{s_{i(n+1)}} \frac{(\pi_{i'j'k'1} - \pi_{i'j'k'3})}{s_{i'(n+1)}}. \end{aligned} \tag{34}$$

Note that (34) is different for the controlled selection procedure than if independent sampling is used to select the sample units in each design.

5. Comparison with Overlap Procedure of Causey, Cox, and Ernst (1985)

Causey, Cox, and Ernst (1985) present an optimal procedure for maximizing overlap of sample units for two designs when the sample units for the two designs are selected sequentially, that is the D_1 sample units are selected first, and then the D_2 sample units are selected with probabilities conditioned on the set of D_1 sample units selected. Their procedure also uses a transportation theory algorithm, although in quite a

different way than used by the controlled selection approach in this article for simultaneous selection. In the Introduction we remarked that simultaneous selection of sample units for the two designs allows for generally higher overlap than sequential selection. To illustrate this point consider the example presented in Section 2. By optimality of our procedure, the maximum overlap for any procedure is 1.9, the expected overlap for the controlled selection procedure. However, if the two D_1 sample units are selected first and also selected independently of each other, then there is a 0.04 probability that B_{121} and B_{221} are the two selected D_1 units. Since these two units are in the same D_2 stratum they cannot both be in the D_2 sample, reducing the maximum overlap when the D_1 units are selected first to 1.86, which is the expected overlap for this example when using the procedure of Causey, Cox, and Ernst (1985). The controlled selection procedure avoids this 0.04 reduction in overlap by not allowing these two units to be in the D_1 sample together. In particular, with the set of arrays in (9), neither unit can be in the D_1 sample if $\tilde{n}_1$ is selected, only B_{221} if $\tilde{n}_2$ or $\tilde{n}_4$ is selected, and only B_{121} if $\tilde{n}_3$ is selected.

6. Minimization of Overlap

Sometimes it is considered desirable to minimize the expected number of sample units in common to two designs rather than maximize it. Reduction of respondent burden is one reason for minimizing overlap. The procedure described in Sections 2 and 3 can very easily be modified to minimize overlap. Simply let $\pi_{ijk4} = \max\{\pi_{ijk1} + \pi_{ijk2} - 1, 0\}$ and substitute π_{ijk4} for π_{ijk3} in (12), (18), (26) and (27). The remainder of the procedure is identical to the maximization procedure.

The rationale for the definition of π_{ijk4} in the minimization case is analogous to the rationale for the definition of π_{ijk3} in the maximization case presented in Section 2. For while π_{ijk3} is the maximum possible value for $P(B_{ijk} \in S_1 \cap S_2)$, the minimum possible value for this probability is π_{ijk4}.

7. Modifications for Other Designs

A key assumption in the procedure presented in Sections 2 and 3 is that both the D_1 and D_2 designs are one unit per stratum. The author is unaware of how to apply this procedure for other designs, unless the design allows for a unit to be selected more than once for the same design.

The first step in the two-step procedure can easily be modified for other designs, including the general case when, in place of (1), $v_{i1}, i = 1, \ldots, m$, and v_{j2}, $j = 1, \ldots, n$, are sets of positive integers with v_{i1} units to be selected from D_1 stratum i and v_{2j} units are selected from D_2 stratum j. However, there is a major difficulty in generalizing the second step which arises from the fact that in the general case, unlike the one unit per stratum case, the sample units corresponding to each internal cell in the array $\tilde{n}$ cannot be selected independently of the sample units in all other internal cells. This is because whenever $v_{i1} > 1$, for example, it may occur that $n_{i(n+1)} \geq 1$ and also $n_{ij} \geq 1$ for some $j = 1, \ldots, n$, and if the selection of sample units is conducted independently from cell to cell, the same unit in the intersection of the ith D_1 and

jth D_2 stratum may be selected twice, from cell ij and cell $i(n+1)$. To avoid selecting the same unit twice, some form of without replacement sampling would be needed, but it is not clear to the author how this can be done in this context.

8. Application to the Proposed Expansion of the Current Population Survey

The proposed, but since canceled, expansion of the CPS could have been an important application of the controlled selection procedure described in the preceding sections. The following is a general outline of this proposal. (For further details see Tupek, Waite, and Cahoon (1990).) Beginning in 1994, a redesign of the CPS (the D_1 design) was being phased in. This design has precision requirements for monthly estimates for the nation and the larger states, and for annual estimates for the remaining states and the District of Columbia. Beginning in 1996, if the proposal had been implemented, a sample expansion (the D_2 design) would have taken place to meet reliability requirements for monthly estimates for all 50 states and the District of Columbia.

The CPS is a multistage stratified design. Four methods for selecting PSUs for the D_1 and D_2 designs are described and compared on the basis of variances in Weidman and Ernst (1991). For the purposes of illustrating the procedure of the current article, we consider two of them, the controlled selection method, and the independent sample method. Both methods select the D_1 and D_2 sample PSUs from the same optimal one PSU per stratum D_1 and D_2 stratifications. Due to the more stringent reliability requirements for D_2 than for D_1, (5) holds and hence each D_1 sample PSU is a D_2 sample PSU under controlled selection.

For the independent sample method, the D_2 sample PSUs are selected independently of the D_1 sample PSUs. For this method we calculated that of the 257 noncertainty D_1 sample PSUs, the expected number retained in the D_2 sample is 174.5 or 67.9%, and thus for the independent sample method, unlike controlled selection, a large number of sample PSUs are not retained. These 257 sample PSUs are selected from a universe of 1140 D_1 noncertainty PSUs. In the D_2 design these 1140 PSUs form 518 strata, 200 of which are certainty strata.

In Weidman and Ernst (1991) it is calculated that the variances for key labor force characteristics for the D_2 design are generally quite similar for controlled selection and independent sample. Thus, at least for this application, controlled selection would retain all the D_1 sample PSUs in the D_2 sample without a variance penalty. However, as previously mentioned, controlled selection is not usable in applications where the D_2 sample PSUs are selected subsequent to the selection of the D_1 sample PSUs. Furthermore, as illustrated by (34), variance estimation would be more complex for controlled selection than for some other approaches to the selection of PSUs.

9. References

Brewer, K.R.W., Early, L.J., and Joyce, S.F. (1972). Selecting Several Samples from a Single Population. Australian Journal of Statistics, 14, 231–239.

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. (1986). Maximizing the Overlap Between Surveys When Information Is Incomplete. European Journal of Operational Research, 27, 192–200.

Ernst, L.R. (1993). Simultaneous Selection of Primary Sampling Units for Two Designs. U. S. Bureau of the Census, Statistical Research Division, Research Report Series, No. RR-90/10.

Glover, F., Karney, D., Klingman, D., and Napier, A. (1974). A Computation Study on Start Procedures, Basic Change Criteria and Solution Algorithms for Transportation Problems. Management Sciences, 20, 789–813.

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.

Kish, L. and Scott, A. (1971). Retaining Units After Changing Strata and Probabilities. Journal of the American Statistical Association, 66, 461–470.

Perkins, W.M. (1970). 1970 CPS Redesign: Proposed Method for Deriving Sample PSU Selection Probabilities Within 1970 NSR Strata. Memorandum to Joseph Waksberg, U.S. Bureau of the Census.

Perry, C.R., Burt, J.C., and Iwig, W.C. (1993). Methods of Selecting Samples in Multiple Surveys to Reduce Respondent Burden. Proceedings of the International Conference on Establishment Surveys, American Statistical Association, 345–351.

Pruhs, K. (1989). The Computational Complexity of Some Survey Overlap Problems. Proceedings of the Section on Survey Research Methods, American Statistical Association, 747–752.

Raj, D. (1968). Sampling Theory. New York: McGraw Hill.

Tupek, A.R., Waite, P.J., and Cahoon, L.S. (1990). Sample Expansion Plans for the Current Population Survey. Proceedings of the Section on Survey Research Methods, American Statistical Association, 72–77.

Weidman, L. and Ernst, L.R. (1991). Multiple Workloads per Stratum Sampling Designs. Proceedings of the Section on Survey Research Methods, American Statistical Association, 443–448.

Received October 1992
Revised March 1996