

April 17, 1995

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

by

Lawrence R. Ernst  
Bureau of Labor Statistics  
Office of Compensation and Working Conditions  
Research Group  
Washington, D.C. 20212

**Abstract:** 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.

**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 paper 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.

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

There is one unit selected from each  $D_1$  and each  $D_2$  stratum. (1.1)

Each unit is selected into each design with the required probability. (1.2)

The expected value for the number of sample units common to the two designs is maximized. (1.3)

In this paper 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. (1.4)

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. 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 paper, 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 paper. Using a graph theory approach, he presented an algorithm which satisfies conditions (1.1) - (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 et al. 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 (1.1) - (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. (1.5)

For this special case it can be shown that (1.3) and (1.4) can be replaced with the more stringent requirement that:

Each  $D_1$  sample unit is a  $D_2$  sample unit. (1.6)

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 Current Population Survey, 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 paper 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 to be in sample for both the  $i$ -th  $D_1$  stratum and the  $j$ -th  $D_2$  stratum, 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 (1.1) and (1.4), and the probability mechanism for selecting the array guarantees that (1.3) 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 (1.2).

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 (1.1), (1.3) and (1.4), but postpone to Section 3 discussion of how the controlled selection algorithm can be applied to insure 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 (1.2). 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  $i = 1, \dots, m$ ,  $j = 1, \dots, n$ ,  $n_{ij}$  is a 0,1 variable which indicates the number of units that are to be selected to be in sample for both the  $i$ -th  $D_1$  stratum and the  $j$ -th  $D_2$  stratum. Since the two designs are in general different, some units may have to be in sample for only one design.  $n_{i(n+1)}$ ,  $i = 1, \dots, m$ , is a 0,1 variable that indicates the number of units in the  $i$ -th  $D_1$  stratum that are to be  $D_1$  only sample units, and similarly  $n_{(m+1)j}$ ,  $j = 1, \dots, n$ , is the number of units in the  $j$ -th  $D_2$  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

$$n_{i(n+2)} = \sum_{j=1}^{n+1} n_{ij}, \quad i = 1, \dots, m+2, \quad (2.1)$$



$$n_{(m+2)j} = \sum_{i=1}^{m+1} n_{ij}, \quad j = 1, \dots, n+2. \quad (2.2)$$

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 the example there are the following four values for  $\tilde{n}$ , denoted  $\tilde{n}_1 - \tilde{n}_4$ :

$$\begin{array}{r}
\tilde{n}_1 = \begin{array}{cccc|c} 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 \\ \hline 1 & 1 & 1 & 0 & 3 \end{array} \\
\tilde{n}_2 = \begin{array}{cccc|c} 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 \\ \hline 1 & 1 & 1 & 0 & 3 \end{array} \\
\tilde{n}_3 = \begin{array}{cccc|c} 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 \\ \hline 1 & 1 & 1 & 0 & 3 \end{array} \\
\tilde{n}_4 = \begin{array}{cccc|c} 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 2 \\ \hline 1 & 1 & 1 & 1 & 4 \end{array}
\end{array} \tag{2.3}$$

The probability that  $\tilde{n}_k$  is selected,  $k=1,2,3,4$ , is denoted  $p_k$ , with .5, .3, .1, .1, respectively, the values of these probabilities for this example. The  $\tilde{n}_k$ 's 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, (2.3) 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  $D_1$  stratum 1 and  $D_2$  stratum 1, and one unit is selected from  $D_1$  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 (1.1), (1.3) and (1.4). To satisfy (1.1) we must have

$$n_{i(n+2)} = 1, \quad i = 1, \dots, m, \tag{2.4}$$

and

$$n_{(m+2)j} = 1, \quad j = 1, \dots, n. \tag{2.5}$$

Note that each of the arrays in (2.3) satisfies these conditions.

Before considering the constraints needed to satisfy (1.3) and (1.4) 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  $i=1, \dots, m$ ,  $j=1, \dots, n$ , let  $t_{ij}$  denote the number of units in the population that are in both the  $i$ -th  $D_1$  stratum and the  $j$ -th  $D_2$  stratum; let  $B_{ijk}$  denote the  $k$ -th such unit,  $k=1, \dots, t_{ij}$ ; and let  $T$  denote the set of all triples  $(i, j, k)$ . For  $(i, j, k) \in T$ , let  $\mathbf{p}_{ijk1}$ ,  $\mathbf{p}_{ijk2}$  denote the preassigned selection probabilities for  $B_{ijk}$  in the  $D_1$  and  $D_2$  designs, respectively, and let  $\mathbf{p}_{ijk3} = \min\{\mathbf{p}_{ijk1}, \mathbf{p}_{ijk2}\}$ . Finally, let

$$s_{ij} = \sum_{k=1}^{t_{ij}} \mathbf{p}_{ijk3} \quad i = 1, \dots, m, \quad j = 1, \dots, n. \quad (2.6)$$

Note that (1.2) is then equivalent to the requirement that

$$P(B_{ijk} \in S_{\mathbf{a}}) = \mathbf{p}_{ijk\mathbf{a}}, \quad (i, j, k) \in T, \quad \mathbf{a} = 1, 2. \quad (2.7)$$

If (2.7) does hold then  $P(B_{ijk} \in S_1 \cap S_2) \leq \mathbf{p}_{ijk3}$  for all  $(i, j, k) \in T$ , which, by (2.6), implies that  $P(n_{ij} = 1) \leq s_{ij}$ ,  $i = 1, \dots, m$ ,  $j = 1, \dots, n$ . Hence, in order to satisfy (1.3) we impose the more restrictive requirement on the random array  $\tilde{n}$  that

$$P(n_{ij} = 1) = s_{ij}, \quad i = 1, \dots, m, \quad j = 1, \dots, n. \quad (2.8)$$

It would follow from (2.8) that the expected number of sample units in common to the two designs is

$$E\left(\sum_{i=1}^m \sum_{j=1}^n n_{ij}\right) = \sum_{i=1}^m \sum_{j=1}^n s_{ij}. \quad (2.9)$$

Consequently, in order to establish (1.4), we impose the further requirement that

$$\left| \sum_{i=1}^m \sum_{j=1}^n n_{ij} - \sum_{i=1}^m \sum_{j=1}^n s_{ij} \right| < 1 \quad (2.10)$$

for each possible value for  $\tilde{n}$ . In Section 3 we demonstrate how to obtain a set of nonnegative integer valued arrays and associated selection probabilities satisfying (2.1), (2.2), (2.4), (2.5), and (2.8-2.10). Below, we demonstrate that these relations together with the method of selecting the units in the second step of the procedure yield (2.7). All of these relations together immediately imply (1.1-1.4).

Also observe that in the special case when (1.5) holds, then  $\mathbf{p}_{ijk3} = \mathbf{p}_{ijk1}$  for all  $(i, j, k) \in T$ . Consequently, by (2.6) the right hand side of (2.9) reduces to  $m$ . However, if the expected number of units in common to the two designs is  $m$  then (1.6) 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 (2.6) and (2.8-2.10). We must first specify values for the  $t_{ij}$ 's,

$p_{iju1}$ 's, and  $p_{iju2}$ 's. In this example,  $D_1$  and  $D_2$  consist of the identical 6 units with the following nonzero  $t_{ij}$ 's:  $t_{11}=2, t_{12}=1, t_{22}=1, t_{23}=2$ . The selection probabilities for each of these 6 units for each design are given in Table 1.

Table 1. Selection Probabilities for Units in Example

	<i>ijk</i>					
	111	112	121	221	231	232
$P_{ijk1}$	.5	.4	.1	.4	.4	.2
$P_{ijk2}$	.4	.6	.3	.7	.6	.4

It can be computed from this table and (2.6), that the  $2 \times 3$  array  $(s_{ij})$  of desired probabilities that  $n_{ij}=1$  is

$$(s_{ij}) = \begin{pmatrix} .8 & .1 & 0 \\ 0 & .4 & .6 \end{pmatrix}; \quad (2.11)$$

that the expected number of units in common to the two designs must be 1.9 units in order to satisfy (2.9); and that there must always be either 1 unit or 2 units in common to satisfy (2.10). It can be calculated that the set of 4 arrays in (2.3) together with their associated probabilities do satisfy (2.8-2.10).

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  $i \leq m, j \leq n$ , for which  $n_{ij}=1$ , we must have  $B_{ijk} \in S_1 \cap S_2$  for a single  $k = 1, \dots, t_{ij}$ . The assigned conditional selection probabilities for these  $t_{ij}$  units are

$$P(B_{ijk} \in S_1 \cap S_2 | n_{ij} = 1) = \mathbf{p}_{ijk3} / s_{ij}, \quad k = 1, \dots, t_{ij}. \quad (2.12)$$

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

$$s_{i(n+1)} = 1 - \sum_{j=1}^n s_{ij}, \quad i = 1, \dots, m, \quad (2.13)$$

$$s_{(m+1)j} = 1 - \sum_{i=1}^m s_{ij}, \quad j = 1, \dots, n, \quad (2.14)$$

$$s_{(m+1)(n+1)} = 0. \quad (2.15)$$

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

$$s_{i(n+2)} = \sum_{j=1}^{n+1} s_{ij}, \quad i = 1, \dots, m+2, \quad (2.16)$$

$$s_{(m+2)j} = \sum_{i=1}^{m+1} s_{ij}, \quad j = 1, \dots, n+2. \quad (2.17)$$

For example, the  $2 \times 3$  array (2.11) expands to the  $4 \times 5$  tabular array

$$\begin{array}{cccc|c} .8 & .1 & 0 & .1 & 1 \\ 0 & .4 & .6 & 0 & 1 \\ .2 & .5 & .4 & 0 & 1.1 \\ \hline 1 & 1 & 1 & .1 & 3.1 \end{array} \quad (2.18)$$

Now (2.1), (2.4), (2.8) and (2.13) imply that the probability that a unit is selected from the  $i$ -th  $D_1$  stratum to be in sample for  $D_1$  only, is

$$P(n_{i(n+1)} = 1) = s_{i(n+1)}, \quad i = 1, \dots, m. \quad (2.19)$$

If  $n_{i(n+1)} = 1$  then  $B_{ijk} \in (S_1 \sim S_2)$  for a single unit among the  $\sum_{j=1}^n t_{ij}$  units in the  $i$ -th  $D_1$  stratum, selected with the assigned conditional probabilities

$$P(B_{ijk} \in (S_1 \sim S_2) \mid n_{i(n+1)} = 1) = (\mathbf{p}_{ijk1} - \mathbf{p}_{ijk3}) / s_{i(n+1)}, \quad j = 1, \dots, n, \quad k = 1, \dots, t_{ij}. \quad (2.20)$$

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

$$P(B_{ijk} \in (S_2 \sim S_1) \mid n_{(m+1)j} = 1) = (\mathbf{p}_{ijk2} - \mathbf{p}_{ijk3}) / s_{(m+1)j}, \quad i = 1, \dots, m, \quad k = 1, \dots, t_{ij}. \quad (2.21)$$

The conditional selection probabilities just defined yield (2.7), since for  $\mathbf{a} = 1$  this follows from (2.8), (2.12), (2.19) and (2.20) by combining

$$P(B_{ijk} \in S_1 \cap S_2) = P(n_{ij} = 1) P(B_{ijk} \in S_1 \cap S_2 \mid n_{ij} = 1) = \mathbf{p}_{ijk3},$$

$$P(B_{ijk} \in (S_1 \sim S_2)) = P(n_{i(n+1)} = 1) P(B_{ijk} \in S_1 \cap S_2 \mid n_{i(n+1)} = 1) = \mathbf{p}_{ijk1} - \mathbf{p}_{ijk3},$$

while (2.7) for  $\mathbf{a} = 2$  is obtained similarly.

To illustrate the second step for the example that we have been considering, suppose that the array  $\tilde{n}_1$  in (2.3) is selected at the first step. Then since  $n_{11} = 1$ , we have by (2.12) 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 (2.21).

### 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 paper, that is to construct a finite set of  $(m+2) \times (n+2)$  nonnegative, integer-valued, tabular arrays,  $\tilde{n}$ , and associated probabilities, satisfying (2.1), (2.2), (2.4), (2.5) and (2.8-2.10).

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 (2.3) is a zero-restricted controlled rounding of (2.18).

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,  $p_1, \dots, p_l$ , satisfying:

$$\tilde{n}_u \text{ is a zero-restricted controlled rounding of } (a_{ij}) \text{ for all } u = 1, \dots, l, \quad (3.1)$$

$$\sum_{u=1}^l p_u = 1, \quad (3.2)$$

$$\sum_{u=1}^l n_{iju} p_u = a_{ij}, \quad i = 1, \dots, m+2, \quad j = 1, \dots, n+2. \quad (3.3)$$

(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  $u$ -th 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  $\tilde{n}_u$ 's with its associated probability, then by (3.1) 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 (3.3) 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 (2.18). A solution to this problem is the set of arrays presented in (2.3), 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  $\tilde{n}_1, \dots, \tilde{n}_l$  and  $p_1, \dots, p_l$ . We proceed to show that with  $(a_{ij}) = (s_{ij})$ , a solution to the controlled selection problem satisfies (2.1), (2.2), (2.4), (2.5) and (2.8-2.10).

(2.1) and (2.2) follow immediately from (3.1). Next observe that it follows from (2.13), (2.14), (2.16) and (2.17) that

$$s_{i(n+2)} = 1, \quad i = 1, \dots, m, \quad \text{and} \quad s_{(m+2)j} = 1, \quad j = 1, \dots, n, \quad (3.4)$$

which together with (3.1) yield (2.4) and (2.5). To obtain (2.8), note that for  $i = 1, \dots, m$ ,  $j = 1, \dots, n$ , we have  $0 \leq s_{ij} \leq 1$  by (2.6), and hence  $n_{iju}$  is a 0,1 variable for all  $u$  by (3.1). (2.8) then follows from (3.3). (2.9) follows immediately from (2.8).

To deduce (2.10) we first obtain from (2.16), (2.15) and (2.14) that

$$s_{(m+1)(n+2)} = \sum_{j=1}^n s_{(m+1)j} = n - \sum_{i=1}^m \sum_{j=1}^n s_{ij}. \quad (3.5)$$

We next note that  $n_{(m+1)(n+1)} = 0$  by (2.15) and (3.1), and then combine this result with (2.1), (2.2) and (2.5) to obtain

$$n_{(m+1)(n+2)} = \sum_{j=1}^n n_{(m+1)j} = \sum_{j=1}^n (n_{(m+2)j} - \sum_{i=1}^m n_{ij}) = n - \sum_{i=1}^m \sum_{j=1}^n n_{ij}. \quad (3.6)$$

Finally, we combine (3.1), (3.5) and (3.6) to conclude

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

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 (3.4), 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 variance formulas are derived for estimators of totals for both designs for the sampling procedure detailed in the previous two sections, assuming single stage sampling. 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 formulas presented here with Raj (1968, p.118).

Let  $X$  denote the total value over the entire population for a characteristic of interest and let  $X_{ijk}$  denote the total for unit  $B_{ijk}$  for each  $(i, j, k) \in T$ . For  $\mathbf{a} = 1, 2$ , let  $\hat{X}_{\mathbf{a}}$  denote the usual estimator for  $X$  for  $D_{\mathbf{a}}$  corresponding to probability proportional to size sampling, that is

$$\hat{X}_{\mathbf{a}} = \sum \frac{X_{ijk}}{p_{ijka}},$$

where the summation is over all  $(i, j, k)$  such that  $B_{ijk} \in S_{\mathbf{a}}$ . For  $(i, j, k), (i^*, j^*, k^*) \in T$ ,  $(i, j, k) \neq (i^*, j^*, k^*)$ ,  $\mathbf{a} = 1, 2$ , let

$$p_{ijki^*j^*k^*\mathbf{a}} = P(B_{ijk}, B_{i^*j^*k^*} \in S_{\mathbf{a}}).$$

Finally, for each  $i, j, i^*, j^*$ , for which  $i \leq m+1$ ,  $i^* \leq m+1$ ,  $j \leq n+1$ ,  $j^* \leq n+1$ , let  $r_{iji^*j^*} = P(n_{iju} = n_{i^*j^*u} = 1)$ , that is the is the sum of  $p_u$  over all  $u$  for which  $n_{iju} = n_{i^*j^*u} = 1$ .

Then from Raj (1968, p. 54),

$$V(\hat{X}_{\mathbf{a}}) = \frac{1}{2} \sum_{\substack{(i,j,k),(i^*,j^*,k^*) \in T \\ (i,j,k) \neq (i^*,j^*,k^*)}} (p_{ijka}p_{i^*j^*k^*\mathbf{a}} - p_{ijki^*j^*k^*\mathbf{a}}) \left( \frac{X_{ijk}}{p_{ijka}} - \frac{X_{i^*j^*k^*}}{p_{i^*j^*k^*\mathbf{a}}} \right)^2. \quad (4.1)$$

Consequently, it is only necessary to show how to compute  $p_{ijki^*j^*k^*\mathbf{a}}$  for each  $(i, j, k), (i^*, j^*, k^*) \in T$ ,  $(i, j, k) \neq (i^*, j^*, k^*)$ . To do this for  $\mathbf{a} = 1$ , first observe that  $p_{ijki^*j^*k^*1} = 0$  if  $i = i^*$ . Therefore, it may be assumed from now on that  $i \neq i^*$ . Then to obtain  $p_{ijki^*j^*k^*1}$ , observe that both  $B_{ijk}$  and  $B_{i^*j^*k^*}$  can be in  $S_1$  if for some  $u$ , either

$$n_{iju} = n_{i^*j^*u} = 1, \quad n_{i(n+1)u} = n_{i^*j^*u} = 1, \quad n_{iju} = n_{i^*(n+1)u} = 1 \quad \text{or} \quad n_{i(n+1)u} = n_{i^*(n+1)u} = 1,$$

which combined with (2.12) and (2.20) yield the four terms in the following expression:

$$\begin{aligned}
\mathbf{p}_{ijki^*j^*k^*1} &= r_{iji^*j^*} \frac{\mathbf{p}_{ijk3}}{s_{ij}} \frac{\mathbf{p}_{i^*j^*k^*3}}{s_{i^*j^*}} \\
&+ r_{i(n+1)i^*j^*} \frac{(\mathbf{p}_{ijk1} - \mathbf{p}_{ijk3})}{s_{i(n+1)}} \frac{\mathbf{p}_{i^*j^*k^*3}}{s_{i^*j^*}} \\
&+ r_{iji^*(n+1)} \frac{\mathbf{p}_{ijk3}}{s_{ij}} \frac{(\mathbf{p}_{i^*j^*k^*1} - \mathbf{p}_{i^*j^*k^*3})}{s_{i^*(n+1)}} \\
&+ r_{i(n+1)i^*(n+1)} \frac{(\mathbf{p}_{ijk1} - \mathbf{p}_{ijk3})}{s_{i(n+1)}} \frac{(\mathbf{p}_{i^*j^*k^*1} - \mathbf{p}_{i^*j^*k^*3})}{s_{i^*(n+1)}}. \tag{4.2}
\end{aligned}$$

The only differences in the expression for  $\mathbf{p}_{ijki^*j^*k^*2}$ , which is obtained similarly, are that the subscripts  $i(n+1)$ ,  $i^*(n+1)$  and 1 are replaced by the subscripts  $(m+1)j$ ,  $(m+1)j^*$  and 2, respectively, and that  $\mathbf{p}_{ijki^*j^*k^*2} = 0$  if  $j = j^*$ .

Note that (4.2), and hence (4.1), are different for the controlled selection procedure than if independent sampling is used to select the sample units in each design. In the latter case,  $\mathbf{p}_{ijki^*j^*k^*a} = \mathbf{p}_{ijk a} \mathbf{p}_{i^*j^*k^*a}$  if either  $a = 1$  and  $i \neq i^*$ , or if  $a = 2$  and  $j \neq j^*$ , and hence there is no between strata component of variance for independent sampling.

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

Causey, Cox and Ernst (1985) presents 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 the controlled selection approach in this paper 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. Since there are only 2 sample units selected for  $D_1$ , the maximum overlap for any procedure can be 2. Furthermore, since  $p_{1111} = .5$  and  $p_{1112} = .4$  there is at least a .1 probability that  $B_{111}$  is in sample for  $D_1$  but not  $D_2$ , reducing the maximum overlap to 1.9, which is exactly 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 .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, further 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 .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 (2.3), 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.

Although this is only one example, the expected overlap when using an overlap procedure with sequential selection can never exceed the expected overlap when using the procedure of this paper for the same reason as illustrated by this example. That is, when the sample units are selected sequentially, the joint probabilities for selecting sets of sample units for  $D_1$  is fixed before applying the overlap procedure, which constrains the set of feasible solutions in comparison with simultaneous selection which allows these probabilities to be determined as part of the optimization process.

## 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  $\mathbf{p}_{ijk4} = \max\{\mathbf{p}_{ijk1} + \mathbf{p}_{ijk2} - 1, 0\}$  and substitute  $\mathbf{p}_{ijk4}$  for  $\mathbf{p}_{ijk3}$  in (2.6), (2.12), (2.20) and (2.21). The remainder of the procedure is identical to the maximization procedure.

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

## 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 (1.1),  $v_{i1}$ ,  $i = 1, \dots, m$ , and  $v_{j2}$ ,  $j = 1, \dots, n$ , are sets of positive integers with  $v_{i1}$  units to be selected from  $D_1$  stratum  $i$  and  $v_{j2}$  units are selected from  $D_2$  stratum  $j$ . However, there is at least one formidable complication in generalizing the second step. We proceed to expand on both of these points.

To generalize the first step, the only changes for the general case in defining the  $(m+2) \times (n+2)$  array  $(s_{ij})$  are that "1" is replaced by " $v_{i1}$ " in (2.13) and by " $v_{i2}$ " in (2.14).

The analogous substitutions are also made in (2.4) and (2.5). The procedure for selecting the array  $\tilde{n}$  is otherwise identical to that for the 1 unit per stratum case. In the general case  $n_{ij}$  is a  $\lfloor s_{ij} \rfloor, \lfloor s_{ij} \rfloor + 1$  rather than a 0,1 variable, with (2.8) consequently replaced by

$$P(n_{ij} = \lfloor s_{ij} \rfloor + 1) = s_{ij} - \lfloor s_{ij} \rfloor, \quad (i, j, k) \in T.$$

The major difficulty in generalizing the second step 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  $j = 1, \dots, n$ , and if the selection of sample units is conducted independently from cell to cell, the same unit in the  $i$ -th  $D_1$  and  $j$ -th  $D_2$  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) is 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 paper, 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  $D_1$ , (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 Ernst and Weidman (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, and that both of the methods produce smaller variances than the other two methods considered, which do not select their  $D_2$  sample PSUs from an optimal  $D_2$  stratification, but which, like controlled selection, insure that the  $D_1$  sample PSUs are a subset of the  $D_2$  sample PSUs. Consequently, if it is required that the  $D_1$  sample PSUs be a subset of the  $D_2$  sample PSUs and that minimization of variance is the chief criteria for choosing

among methods meeting this requirement, then controlled selection is the method of choice for this application among those considered. 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 the variance formulas derived in Section 4, 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
- Chandhok, P., Weinstein, R., and Gunlicks, C. (1990). Augmenting a Sample to Satisfy Subpopulation Requirements. *Proceeding of the Section on Survey Research Methods, American Statistical Association*, 696-701.
- 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 Associations, 443-448.