8/3/95

MAXIMIZING AND MINIMIZING OVERLAP OF ULTIMATE SAMPLING UNITS

Lawrence R. Ernst
Bureau of Labor Statistics, 2 Massachusetts Ave., NE, Room 3160, Washington DC 20212

1. INTRODUCTION

There exists a fairly extensive set of literature, beginning with Keyfitz (1951), on the problem of maximizing the overlap of sampling units retained in sample when redesigning a survey for which the units are selected with probability proportional to size. Procedures for maximizing overlap do not alter the unconditional probability of selection for a unit in the new sample, but condition its probability of selection on the initial sample in such a manner that it is generally greater than its unconditional probability when the unit was in the initial sample and less otherwise. The more recent approaches to the overlap problem which employ linear programming, such as Causey, Cox and Ernst (1985) and Ernst (1986), are also applicable to the analogous problem of minimizing the overlap of sampling units.

Most of the previous work in this area has focused on the overlap of primary sampling units (PSUs) in a multistage stratified design, with each stratum in the new design representing a separate overlap problem. Typically, the motivation for maximizing the overlap of PSUs is to reduce additional costs, such as the training of a new interviewer for a household survey, incurred with each change of PSU. Generally, the number of sample PSUs per stratum is quite small, commonly either one or two, and most of the overlap procedures have the drawback that they are usable only in such situations. In fact, the earlier procedures that do not use linear programming, such as Keyfitz (1951), are not applicable at all to other than one PSU per stratum designs. The linear programming procedures are at least in theory applicable to very general designs, but the size of the linear programming problem commonly increases so rapidly as the number of sample PSUs per stratum increases, that these procedures generally cannot be operationally used for designs with other than a very small number of sample PSUs per stratum.

Overlap procedures have also been used at the ultimate sampling unit (USU) level. For example, Brick, Morganstein and Wolter (1987) describe an application of overlap maximization to selection of post offices for a survey conducted for the Postal Service. The Bureau of Labor Statistics (BLS) uses such a procedure to select establishments from strata for their Occupational Compensation Survey Program (OCSP) (Gilliland 1984). In both applications the number of units selected per stratum is too large for linear programming to be

a viable option. Brick, Morganstein and Wolter (1987) employ a Poisson sampling type of procedure to perform the overlap. Their procedure is quite simple and optimal. However, as is characteristic of Poisson sampling, the procedure does not guarantee a fixed sample size. The procedure used in OCSP is also simple and does guarantee a fixed sample size. However, the procedure is not optimal and is only applicable when within stratum sampling is done with equal probability. Furthermore, in certain circumstances the method produces conditional probabilities greater than 1, with no provision for adjusting for this situation in an unbiased way, that is without altering the new unconditional selection probabilities.

In this paper we present a new overlap procedure with the following properties. It is computationally efficient and hence is usable even when a large number of units are to be selected per stratum. It guarantees a fixed sample size. It is applicable whether the units in a stratum are selected with equal or unequal probabilities. The procedure is unbiased in all situations, that is the unconditional selection probabilities for all units in the new design are preserved. It can be used whether it is desired to maximize overlap or minimize overlap. It can even be used to increase overlap for some units in a stratum, decrease overlap for other units, and to treat a third set of units in a neutral fashion. On the other hand, the procedure is not optimal.

The procedure is presented in Section 2. Actually two procedures are presented. The first procedure, which we call the combined initial strata (CIS) procedure, computes the conditional selection probabilities for all units in a stratum in the new design together. The second procedure, the separate initial strata (SIS) procedure, is a simple modification of (CIS) in which the units in each new design stratum are partitioned into substrata, each consisting of all units from the same initial design stratum, with the conditional probabilities computed separately for each substratum. Neither procedure always produces a superior overlap to the other, although there are advantages to each which are discussed later, in Section 3. In addition, in Section 2 we demonstrate that when equal probability sampling is used in each stratum, and overlap maximization is desired, then SIS reduces to the current overlap procedure used for selecting establishments in OCSP with two modifications, one which in all cases either increases the expected overlap or leaves it unchanged, while the other avoids the problem of producing conditional probabilities greater than 1.

In Section 3 we present several examples to illustrate various aspects of the two procedures. We also further discuss, very briefly, some of the overlap procedures mentioned already and some additional overlap procedures. Finally, in the Appendix we provide proofs of the results of Section 2.

2. THE CIS AND SIS PROCEDURES

We begin with two, single stage stratified designs, with the sample units in each stratum selected with probability proportional to size. The designs will be referred to as the initial design and the new design. The universe of units for the two designs must have some,

although not necessarily all, units in common. The sample units in the initial design have previously been chosen and we wish to select the sample units for each stratum in the new design with probabilities conditional on the set of sample units in the initial design. Let A denote a stratum of noncertainty units in the new design, and let B_1, B_2, B_3 denote the subsets of A consisting of units with which we wish to maximize overlap, minimize overlap, and treat in a neutral fashion, respectively, with respect to the initial sample. Include in B_3 any units that were certainty units in the initial design. Also include in B_3 any units that were not in the universe of units for the initial design, that is "birth units." For each unit in B_3 we simply set its probability of selection in the new design conditional on the initial sample to be equal to its unconditional new selection probability. For the units, denoted $A_1,\ldots,A_M$, in $B_1 \cup B_2$ we proceed to develop conditional selection probabilities for the new design as follows.

Let $p_i', p_i,\ i = 1,\ldots,M$, denote the initial and new selection probabilities for A_i, respectively. Let a denote the random subset of A consisting of all units that were sample units in the initial design. Let $S = \{1,\ldots,M\}$, and let s denote the random subset of S consisting of those integers i for which we prefer to select A_i in the new sample given a, that is those i for which $A_i \in (B_1 \cap a) \cup (B_2 \sim a)$. Let $p_i,\ i = 1,\ldots,M$, denote the probability that $i \in s$, that is $p_i = p_i'$ if $A_i \in B_1$, and $p_i = 1 - p_i'$ if $A_i \in B_2$. Note that for each a there exists a corresponding s. Let C denote the set of all such s.

We seek a set of probabilities $p_{is},\ i = 1,\ldots,M$, for selecting A_i in the new sample conditional on the random set s, satisfying the following conditions:

$$\sum_{i \in s} p_{is} > \sum_{i \in s} p_i \quad \text{if } \varnothing \neq s \neq S, \tag{2.1}$$

$$\sum_{i \in S \sim s} p_{is} < \sum_{i \in S \sim s} p_i \quad \text{if } \varnothing \neq s \neq S, \tag{2.2}$$

$$\sum_{i \in S} p_{is} = \sum_{i \in S} p_i, \tag{2.3}$$

$$E(p_{is}) = p_i, \quad i \in S, \tag{2.4}$$

where the expectation in (2.4) is over all $s \in C$. Conditions (2.1), (2.2) arise from the goal of selecting as many units A_i as possible in the new sample for which $i \in s$. (2.3) is required since a fixed number of units is to be selected for the new sample from I. Finally (2.4) is simply a restatement of the requirement that the overlap procedure must preserve the unconditional selection probabilities in the new design. Note that (2.2) is a redundant condition since it immediately follows from (2.1) and (2.3).

In Section 2.1 the basic CIS procedure is presented without the modifications necessary to insure that no conditional probabilities are greater than 1. (It is assumed, however, even in this subsection that all conditional probabilities are nonnegative.) In Section 2.2 these modifications are presented. In Section 2.3 it is explained how CIS can be easily altered to obtain SIS. Finally, in Section 2.4 the special case of SIS for equal probability sampling within a stratum is presented.

2.1 The Basic CIS Procedure

To meet objectives (2.1-2.4) we proceed as follows. For each $i \in S$ we associate a positive number a_i. As part of the process of obtaining p_{is} we add a_i to p_i only for those $i \in s$. In order to satisfy (2.3) we compensate for this increase by subtracting an amount $b_s p_i$ from each $i \in S$ where, as indicated by the notation, b_s depends on s, but not i. Thus p_{is} takes the form

$$p_{is} = p_i + I_{is} a_i - b_s p_i, \tag{2.5}$$

where $I_{is} = 1$ if $i \in s$ and $I_{is} = 0$ if $i \notin s$.

To determine appropriate values for a_i and b_s we first observe that if (2.4) holds, then by (2.5),

$$E(p_{is}) = p_i + p_i a_i - E(b_s) p_i = p_i,$$

and consequently, abbreviating $d = E(b_s)$, we would then have

$$a_i = d \frac{p_i}{p_i}, \quad i \in S. \tag{2.6}$$

We seek the largest possible value of d in order to obtain a large value for (2.5) when $i \in s$. Now by (2.5), in order for p_{is} to always be nonnegative we must have $b_s \le 1$ for all $s \in C$, which combined with requirements (2.3), (2.5), (2.6) yields

$$d \sum_{i \in s} \frac{p_i}{p_i} - \sum_{i \in S} p_i \le \sum_{i \in s} a_i - b_s \sum_{i \in S} p_i = 0. \tag{2.7}$$

Now the largest possible d for which the left hand side of the inequality in (2.7) does not exceed 0 for any $s \in C$ is

$$\frac{\sum_{i\in S} p_i}{\max_{s\in C} \sum_{i\in s} \frac{p_i}{p_i}}.$$

However, the denominator of this last expression is not generally, readily computable. Instead we compute, as will be explained shortly, an upper bound, denoted u_C, for the denominator and let

$$d = \frac{\sum_{i\in S} p_i}{u_C}.$$

We then combine this relation and (2.6) to obtain

$$a_i = \frac{p_i \sum_{j\in S} p_j}{p_i u_C}, \quad i \in S. \tag{2.8}$$

Finally, we obtain b_s from the equality relationship in (2.7) and (2.8), that is

$$b_s = \frac{\sum_{i\in s} \frac{p_i}{p_i}}{u_C}. \tag{2.9}$$

It is established in the Appendix that with the specified values for a_i and b_s, (2.5) does satisfy (2.1-2.4).

To compute u_C, the remaining step in the procedure, proceed as follows. Let I_t, $t = 1,\ldots,N$, denote the initial strata that intersect A, and let N_t denote the number of units in I_t. Let n_{t1} denote the sample size for I_t and let $n_{t2} = N_t - n_{t1}$. For $t = 1,\ldots,N$, $j = 1,2$, let M_{tj} denote the number of elements in $I_t \cap B_j$ and $m'_{tj} = \min\{n_{tj}, M_{tj}\}$. Let s' be a subset of S of size $\sum_{t=1}^{N}\sum_{j=1}^{2} m'_{tj}$ such that s' consists of m'_{tj} elements i for each t,j, with these elements corresponding to m'_{tj} units A_i in $I_t \cap B_j$ with the largest values of p_i / p_i. Then let

$$u_C = \sum_{i\in s'} \frac{p_i}{p_i}. \tag{2.10}$$

Clearly $u_C \geq \max_{s \in C} \sum_{i \in s} \frac{p_i}{p_i}$. The reason that equality may not hold is that it is possible that $s' \notin C$. For example, if $m'_{t1} = 2$ for some *t*, and systematic sampling was used to select the units in the initial sample, then the two units in $I_t \cap B_1$ with the largest values of p_i / p_i might never be in the initial sample together, in which case $s' \notin C$.

For use in the next subsection, we let $D_i = C \cap \{s\colon\ i \in s\}$, $i \in S$, and analogous to (2.10), proceed to explain how to compute a lower bound l_{D_i} on $\min_{s \in D_i} \sum_{j \in s} \frac{p_j}{p_j}$, from which it would follow that

$$l_{D_i} / u_C \leq b_s \quad \text{if } i \in s \in C\,. \tag{2.11}$$

For $t = 1,\ldots,N$, let $m''_{t1} = \max\{M_{t1} - n_{t2}, 0\}$, $m''_{t2} = \max\{M_{t2} - n_{t1}, 0\}$. For $i \in S$, $j = 1,2,$ let $m''_{itj} = \max\{m''_{tj}, 1\}$ if $A_i \in I_t \cap B_j$, and $m''_{itj} = m''_{tj}$ otherwise. Then for $i \in S$, let s''_i be a subset of *S* of size $\sum_{t=1}^{N} \sum_{j=1}^{2} m''_{itj}$ such that for each *t,j* for which $A_i \notin I_t \cap B_j$ there are m''_{tj} elements *k* in s''_i, corresponding to m''_{tj} units A_k in $I_t \cap B_j$ with the smallest values of p_k / p_k; while for the *t,j* for which $A_i \in I_t \cap B_j$, we have $i \in s''_i$ in addition to $m''_{itj} - 1$ elements *k* corresponding to units A_k in $I_t \cap B_j \sim \{A_i\}$ with the smallest values of p_k / p_k. Then let

$$l_{D_i} = \sum_{k \in s''_i} \frac{p_k}{p_k}. \tag{2.12}$$

2.2 Modification of CIS to Avoid Conditional Probabilities Greater than 1

The procedure described above requires modification if $p_i + a_i - b_s p_i > 1$ for any *i*, *s*, with $i \in s$, since then $p_{is} > 1$ by (2.5). To avoid obtaining a conditional probability above 1 we proceed to define p_{is} $S_1 = S$, $_1 = C$, and $_{i1} = p_i$.

$k \geq 1$ let b_{sk}, D_{ik}, $'$ r_{ik}, r_k, $+1)$ $C_{(k+1}$, $p_{i(\ +\ )}$ be defined as follows. a_{ik} for i $_k$ b_{sk} are obtained by substituting a_{ik}, b_{sk}, $s \cap S_k$, S_k, C_k, p'_{ik} for a_i, b_s, s, S, C, p_i in (2.8) and (2.9). Then define $D_{ik} = C_k \cap \{s\colon\ i \in s\}$ and let:

$$b'_{ik} = \frac{l_{D_{ik}}}{u_{C_k}}, \quad i \in S_k, \tag{2.13}$$

$$r_{ik} = \frac{1 - p_i - \sum_{j=1}^{k-1} (a_{ij} - b'_{ik} p'_{ij}) r_j}{a_{ik} - b'_{ik} p'_{ik}}, \quad i \in S_k, \tag{2.14}$$

where the summation in (2.14) is understood to be 0 for $k = 1$,

$$r_k = \min\{\min\{r_{ik}: \ i \in S_k\}, 1\}, \tag{2.15}$$

$$S_{k+1} = S_k \sim \{i: \ r_{ik} = r_k\}, \tag{2.16}$$

$$C_{k+1} = \{s \cap S_{k+1}: \ s \in C\}, \tag{2.17}$$

$$p'_{i(k+1)} = (1 - r_k) p'_{ik}, \quad i \in S_{k+1}. \tag{2.18}$$

Note that in defining u_{C_k}, $l_{D_{ik}}$, replace S, B_j in the definitions of u_C, l_{D_i} in Section 2.1 by $S_k, B_j \cap S_k$, respectively.

Finally, let

$$k' = \min\{k: \ r_k = 1 \text{ or } S_{k+1} = \varnothing\}, \tag{2.19}$$

$$k_i = \max\{k: \ i \in S_k, \ k \le k'\}, \ \ i \in S, \tag{2.20}$$

$$p_{is} = p_i + \sum_{k=1}^{k_i} r_k (l_{is} a_{ik} - b_{sk} p'_{ik}). \tag{2.21}$$

The general idea of this iterative procedure is that the definitions of r_{ik}, r_k together with the relation

$$b'_{ik} \le b_{sk} \text{ for } i \in s \cap S_k, \tag{2.22}$$

(which follows from (2.11), (2.13)) keep the p_{is} defined in (2.21) from getting above 1; while (2.18) is used to insure that $p_{is} \ge 0$. More details are provided in the Appendix where it is proven that $0 \le p_{is} \le 1$ for all i, s and that (2.1)-(2.4) hold.

Also note that if $r_1 = 1$ then (2.21) reduces to (2.5).

2.3 The SIS Procedure

SIS is an alternative to the CIS procedure, defined in the previous two subsections, for which the conditional probabilities of selection in the new design for units in $I_t \cap (B_1 \cup B_2)$ are computed separately for each *t*, instead of being computed for all units in $B_1 \cup B_2$ together. That is, the conditional probabilities are computed using (2.5-2.21) but with $S \cap \{i\colon A_i \in I_t\}$ and $s \cap \{i\colon A_i \in I_t\}$ replacing *S* and *s*, respectively. Neither CIS nor SIS always yields a larger overlap than the other. However, each approach has a specific advantage over the other that will be discussed and illustrated in the example in Section 3.1.

2.4 SIS with Equal Probability Sampling

For overlap maximization, there is one situation where SIS provides a particularly simple set of conditional probabilities, that is the case when equal probability sampling is used within each initial and new stratum. We let *m* denote the sample size for *A* in the new design; m_t denote the number of elements in *s* which correspond to elements in I_t; drop the subscript "1" from n_{t1}, M_{t1} defined in Section 2.2; and replace the subscript "1" in r_1 with the subscript *t* to denote dependence on I_t. Then, as proven in the Appendix, we have that $k' = 1$, and for each $i \in S$ for which $A_i \in I_t$,

$$\begin{aligned} p_{is} &= \frac{m}{M}\left(1 + \frac{r_t(M_t - m_t)}{\min\{n_t, M_t\}}\right) \text{ if } i \in s, \\ &= \frac{m}{M}\left(1 - \frac{r_t m_t}{\min\{n_t, M_t\}}\right) \text{ if } i \notin s, \end{aligned} \tag{2.23}$$

where

$$r_t = \min\left\{\frac{(M-m)\min\{n_t, M_t\}}{m(M_t - \max\{M_t - N_t + n_t, 1\})}, 1\right\}. \tag{2.24}$$

The analogous formulas for the equal probability case for CIS are not presented, since we do not always have that $k' = 1$ for CIS, and consequently CIS does not produce as simple a formulation as SIS in this case.

The conditional selection probabilities defined by (2.23) and (2.24) differ from those currently used in OCSP in two ways. The current OCSP overlap procedure uses n_t instead of $\min\{n_t, M_t\}$ in the two places it appears in (2.23). When these two values differ, $\min\{n_t, M_t\}$ produces a higher conditional expected overlap. In addition, the current OCSP procedure always sets $r_t = 1$, which can result in conditional probabilities greater than 1, a problem avoided by (2.24).

OCSP sample selection also involves some birth units, that is establishments that were not in existence at the time the initial sample was chosen. The current OCSP overlap procedure samples birth units in the same way that we would, by considering them to be elements of B_3.

3. EXAMPLES

In this section we consider three examples. The first example is an overlap maximization problem, in which the modifications of Section 2.2 to avoid conditional probabilities greater than 1 are not needed. The second example is an overlap minimization problem using the same data as the first example, which also does not require the modifications of Section 2.2. The third example is an overlap maximization problem in which the modifications to avoid conditional probabilities greater than 1 are required.

3.1 Example 1

In this example, using the notation of Section 2, $M = 5$, $N = 2$, with $I_1 \cap A = \{A_1, A_2, A_3\}$, $I_2 \cap A = \{A_4, A_5\}$. We wish to maximize overlap with each of these units, that is $B_1 = A$ and $p_i = p_i'$ for all i. One unit had been selected from each of I_1, I_2 for the initial sample and one unit is to be selected from A for the new sample. The initial and new selection probabilities for each unit are given in Table 1.

Table 1. Selection Probabilities for Units in Example 1.

	i				
	1	2	3	4	5
p_i	.1	.2	.2	.3	.1
p_i	.1	.26	.18	.36	.1
$a_i = b_{i1}'$	.4	.52	.36	.48	.4
r_{i1}	2.5	1.923	2.778	2.083	2.5
$p_{i\{3,4\}}$	.016	.042	.389	.538	.016
a_i^*	.415	.54	.374	.46.	.383
$b_{i1}^{'*}$	.769	1	.692	1	.833
r_{i1}^*	2.659	2.643	3.29	6.4	3
$p_{i\{3,4\}}^*$	.031	.08	.429	.46	0
$p_{i\{3\}}$	.064	.166	.475	.23	.064
$p_{i\{3\}}^*$	.031	.08	.429	.36	.1

We first proceed to compute p_{is} for CIS for each i when $a = \{A_3, A_4\}$ and hence $s = \{3, 4\}$. We first compute u_C. We have $M_{11} = 3$, $M_{21} = 2$, $M_{12} = M_{22} = 0$, $n_{11} = n_{21} = 1$, and

consequently $m'_{11} = m'_{21} = 1,\ m'_{12} = m'_{22} = 0$. Therefore, $s' = \{2,4\}$ and then $u_C = 2.5$ by (2.10). Then, from (2.8) and (2.9) we obtain $b_{\{3,4\}} = .84$, and the set of a_i's in Table 1. To compute b'_{i1}, observe that since the sum of the p_i's is less than 1 for the first 3 units and for the last 2 units, then $n_{t2} \geq M_{t1},\ t = 1,2,$. Consequently, $m''_{tj} = 0$ for all t, j. Therefore, $s''_i = \{i\}$ for all i, and hence $b'_{i1} = a_i$ by (2.8), (2.13). We then obtain the set of r_{i1}'s in Table 1 from (2.14). Since $r_1 = 1$ by (2.15), the set of $p_{i\{3,4\}}$'s in Table 1 can be computed from (2.5).

Note that for this example, the probability of overlap conditional on $s = \{3,4\}$ using CIS is $p_{3\{3,4\}} + p_{4\{3,4\}} = .926$, in comparison with an overlap probability of $p_3 + p_4 = .54$ if the new units are selected independently of the initial sample units.

For the same example the conditional probabilities were also computed for SIS. The values of the corresponding variables, which are indicated by *'s to distinguish them from the variables in CIS, are given in the four rows following the $p_{i\{3,4\}}$ row. We also have that $u^*_{1C} = 1.3$, $u^*_{2C} = 1.2$; $b^*_{1\{3,4\}} = .692$ and $b^*_{2\{3,4\}} = 1$, where the first subscript in each of these variables indicates the initial stratum number. The conditional probability of overlap for SIS is .889, which is less than that for CIS.

For other pairs SIS produces a higher overlap than CIS for this example. To illustrate, if $s = \{3,5\}$ then the conditional probability of overlap is .827 for CIS and .829 for SIS. However, for each of the five singleton sets we have a higher overlap probability for CIS. This is illustrated for the case $s = \{3\}$ by the last two rows in Table 1, since $p_{3\{3\}} = .475$ and $p^*_{3\{3\}} = .429$. A key reason for the higher value of $p_{3\{3\}}$ is that $p^*_{i\{3\}} = p_i > p_{i\{3\}}$ for $i \in \{4,5\}$. The equality part of this relationship occurs because as a result of SIS computing the conditional probabilities separately for each I_t, it does not take into account that $3 \in s$ in the computation of the conditional probabilities for units in I_2, a shortcoming not shared by CIS.

For this example, the unconditional probability of overlap, that is the expected value of the conditional overlap probability over all initial samples, is higher for CIS (.473) than for SIS (.416). Independent selection of the new units in comparison yields an unconditional overlap probability of .216.

These two procedures do not require that the initial sample units in I_1 and I_2 be selected independently of each other. (In fact, as explained in Ernst (1986), previous use of an overlap procedure generally destroys stratum to stratum independence.) However, if this independence assumption does hold then the unconditional probability of overlap can be computed for any overlap procedure. For this example, the simple procedure due to Perkins (1970), which is limited to one unit per stratum designs, has an overlap probability of .443. This is the procedure used by the Census Bureau in the 1970s in redesigning the household surveys that they conduct, and which is still used by BLS for PSU selection for the Consumer

Expenditure Survey. (Perkins' procedure is a generalization of Keyfitz's procedure to the case when the stratifications may be different in the initial and new designs.)

The optimal transportation problem (a form of linear programming) procedure of Causey, Cox and Ernst (1985) has an overlap probability of .7. This is clearly optimal for this example since it is precisely the probability that at least one of the units in *A* was in the initial sample. This procedure can result in very large transportation problems, even for small number of units per stratum, and this author is unaware of its use in sample selection for any survey. A modified version of this procedure (Ernst and Ikeda 1994), which can result in dramatically smaller transportation problems, also yields an overlap of .7, although it does not always produce an optimal overlap. This procedure was used in PSU selection for the 1990s' redesign of the Census Bureau's Survey of Income and Program Participation. These last two procedures, unlike the other procedures, require that the sample units in the initial sample were selected independently from stratum to stratum. (The other procedures mentioned in this section require this assumption only to be able to compute the overlap probability, while these two procedures require it even to be able to meet the condition of preserving unconditional selection probabilities in the new design.) A third linear programming procedure (Ernst 1986), for use when this independence requirement is not met has been used by the Census Bureau in the redesign of several household surveys in the 1980s and 1990s. It yields an overlap of .61 for this example.

Although CIS produces a higher unconditional probability of overlap than SIS for this example, this is not always the case. In fact, consider the same example with the only modifications that $p_4 = p_4' = .1$ and $p_5 = p_5' = .03$. In this case, the expected overlap is .277 for CIS and .297 for SIS. The reason that these changes reduce the expected overlap more for CIS than SIS is that they result in much larger values for $p_i / p_i,\ i = 4,5$. For CIS this results in a larger value of u_C and hence a smaller value of a_i for $i = 1,2,3$. However, for SIS the change affects only the value of u_{2C}^*, not u_{1C}^*, and hence leaves $a_i^*,\ i = 1,2,3$, unchanged. Thus it appears that SIS may yield a higher overlap when the values of p_i / p_i vary widely across initial strata.

3.2 Example 2

The second example, presented in Table 2, only differs from the first in that we now wish to minimize overlap with all units in *A* that were in the initial sample, that is $A = B_2$. Then $p_i = 1 - p_i'$ for all *i*. As in the first example, we first compute p_{is} for $a = \{A_3, A_4\}$, except we now have $s = \{1,2,5\}$ To compute u_C, we note that now $m_{t1}' = M_{t1} = 0,\ t = 1,2$, $M_{12} = 3 \le n_{12},\ M_{22} = 2 \le n_{22}$. Consequently, $m_{12}' = 3,\ m_{22}' = 2$, and hence $s' = S$, $u_c = 1.287$. We then readily obtain $b_{\{1,2,5\}} = .425$ and the set of a_i's in Table 2. To compute the b_{i1}''s, we note that $m_{11}'' = m_{21}'' = 0,\ m_{12}'' = 2,\ m_{22}'' = 1$, with the resulting s_i'''s , b_{i1}''s and r_{i1}'s presented in Table 2. The $p_{i\{1,2,5\}}$'s are then readily computable from (2.5) since $r_1 = 1$.

For SIS the calculations are similar, with only the $p^*_{i\{1,2,5\}}$'s presented in Table 2. The conditional overlap is .310 for CIS, .357 for SIS, and, as in Example 1, .540 for independent selection.

Table 2. Selection Probabilities for Units in Example 2.

	i				
	1	2	3	4	5
$p_i = 1 - p_i'$	.9	.8	.8	.7	.9
p_i	.1	.26	.18	.36	.1
a_i	.086	.253	.175	.4	.086
s_i''	{1,3,5}	{1,2,5}	{1,3,5}	{1,3,4}	{1,3,5}
b_{i1}'	.348	.425	.348	.661	.348
r_{i1}	17.441	5.210	7.301	3.956	17.441
$p_{i\{1,2,5\}}$	.144	.402	.103	.207	.144
$p^*_{i\{1,2,5\}}$	.125	.354	.061	.296	.164

3.3 Example 3

The assumptions of this example differ from those of the first in only one way. Two units are now selected from *A* for the new design, doubling the values of the p_i's. The computations for $s = \{3,4\}$, which are presented for CIS only, proceed similarly to the first example with $u_{C_1} = u_C = 5$, $b_{\{3,4\}1} = b_{\{3,4\}} = .84$ and the a_{i1}'s, $b_{i1}^{'}$'s, r_{i1}'s given in Table 3. Now, however, since $r_1 = .456$, we must use the recursive process (2.13)-(2.21). We have $S_2 = \{1,2,3,5\}$, and the p_{i2}''s, obtained from (2.18), given in Table 3. Then proceeding as in the first step, with S, C, p_i replaced by S_2, C_2, p_{i2}, we obtain $u_{C_2} = 2.504$, $b_{\{3,4\}2} = .391$, and the indicated a_{i2}'s, b_{i2}''s, r_{i2}'s. Since $r_2 = .553$, the process continues, with $S_3 = \{1,3,5\}$, $u_{C_3} = .973$, $b_{\{3,4\}3} = .45$, and the other values corresponding to $k = 3$ in Table 3. Now $r_3 = 1$, $k' = 3$, and the process stops with the p_{is}'s obtained from (2.21). The conditional expected overlap is 1.584 for CIS and 1.08 for independent selection.

Table 3. Selection Probabilities for Units in Example 3

	i				
	1	2	3	4	5
p_i	.1	.2	.2	.3	.1
$p'_{i1} = p_i$	.2	.52	.36	.72	.2
$a_{i1} = a_i$	.8	1.04	.72	.96	.8
b'_{i1}	.4	.52	.36	.48	.4
r_{i1}	1.111	.624	1.084	.456	1.111
p'_{i2}	.109	.283	.196		.109
a_{i2}	.303	.394	.273		.303
b'_{i2}	.435	.565	.391		.435
r_{i2}	1.846	.553	1.893		1.846
p'_{i3}	.049		.088		.049
a_{i3}	.092		.083		.092
b'_{i3}	.5		.45		.5
r_{i3}	4.852		5.995		4.852
$p_{i\{3,4\}}$	.078	.26	.702	.882	.078

APPENDIX

Proof that (2.5) satisfies (2.1-2.4). To establish (2.1), we combine (2.8) and (2.9), obtaining

$$\sum_{i \in s} (a_i - b_s p_i) = \frac{\sum_{i \in s} \frac{p_i}{p_i}}{u_C} \left(\sum_{i \in S} p_i - \sum_{i \in s} p_i \right) > 0 , \tag{A.1}$$

which we combine with (2.5) to conclude (2.1).

We obtain (2.3) similarly, since

$$\sum_{i \in s} a_i - \sum_{i \in S} b_s p_i = \frac{\sum_{i \in s} \frac{p_i}{p_i}}{u_C} \left(\sum_{i \in S} p_i - \sum_{i \in S} p_i \right) = 0 . \tag{A.2}$$

(2.2) follows immediately from (2.1) and (2.3)

Finally, to establish (2.4), observe that by (2.8) and (2.9), we have

$$p_i a_i = E(b_s)p_i, \tag{A.3}$$

which combined with (2.5) yields (2.4)

Proof that p_{is} *defined by (2.21) satisfies* $0 \le p_{is} \le 1$. To establish that $p_{is} \le 1$, where p_{is} is as defined in (2.21), first let

$$f_{ik} = p_i + \sum_{j=1}^{k-1} r_j (a_{ij} - b'_{ij} p'_{ij}),\ k = 1,\ldots,k_i+1;\ \ g_{ik} = a_{ik} - b'_{ik} p'_{ik},\ k = 1,\ldots,k_i.$$

We proceed to establish by induction on k that for $k = 1,\ldots,k_i,\ \ f_{ik} < 1,\ p_{ik} > 0$, $g_{ik} > 0$, and $r_{jk} > 0$ for $j \in S_k$ (and hence

$$r_k > 0 \tag{A.4}$$

by (2.15)). We also show that $f_{i(k_i+1)} \le 1$. It will then follow from these relations and (2.21), (2.22), that $p_{is} \le f_{i(k_i+1)} \le 1$ for $i \in s$, and $p_{is} \le p_i < 1$ for $i \notin s$. Now for $k = 1$ we have $f_{i1} = p_{i1} = p_i < 1$. To establish that $g_{i1} > 0$ we observe that by (A.3),

$$p_i a_{i1} = E(b_{s1})p'_{i1} = p_i E(b_{s1} \mid i \in s)p'_{i1} + (1-p_i)E(b_{s1} \mid i \notin s)p'_{i1},$$

and hence by (2.22), $a_{i1} > E(b_{s1} \mid i \in s)p'_{i1} \ge b'_{i1} p'_{i1}$. Finally, for $j \in S$ we have by (2.14) that $r_{j1} = (1-p_j)/g_{j1} > 0$.

To prove that if the indicated relations hold for $k < k_i$ then they hold for $k+1$, observe that by (2.14),

$$f_{i(k+1)} = f_{ik} + r_k g_{ik} < f_{ik} + r_{ik} g_{ik} = 1, \tag{A.5}$$

(where the inequality in (A.5) is strict by (2.20), (2.16)); $p'_{i(k+1)} > 0$ by (2.18) and fact that $r_k < 1$ for $k < k'$, by (2.19); the relation $g_{i(k+1)} > 0$ is established in the same manner as $g_{i1} > 0$; and $r_{j(k+1)} = (1 - f_{j(k+1)})/g_{j(k+1)} > 0$ for $j \in S_{(k+1)}$. Finally $f_{k_i+1} \le 1$ follows from (A.5), except "<" is replaced by "≤".

To establish that $p_{is} \geq 0$, first note that from (A.1) with $a_{ik}, b_{sk}, p_{ik}, s \cap S_k$ substituted for a_i, b_s, p_i, s, it follows that

$$\sum_{i \in s \cap S_k} (a_{ik} - b_{sk} p'_{ik}) \geq 0, \quad k = 1, \ldots, k_i, \tag{A.6}$$

and that $b_{sk} \leq 1$ for all s, k by (2.9). We then combine these inequalities with (2.21), (2.18), (2.15), (A.4) to conclude that

$$p_{is} \geq p_i - \sum_{k=1}^{k_i} r_k p'_{ik} = p_i \left(1 - \sum_{k=1}^{k_i} r_k \prod_{j=1}^{k-1} (1 - r_j) \right) = p_i \prod_{k=1}^{k_i} (1 - r_k) \geq 0, \tag{A.7}$$

where the last equality in (A.7) can be established by substituting v for k_i in this equation and then proving by induction on v that the equation does hold for $v = 1, \ldots, k_i$.

Proof that (2.21) satisfies (2.1-2.4). Note that (A.6) is a strict inequality by (A.1) if $k = 1$, since $\varnothing \neq s = s \cap S_1 \neq S$. This observation, (2.21), (A.4), (A.6) yield (2.1).

Similarly (2.3) follows from (A.2), with the appropriate substitutions, and (2.21); while (2.1) and (2.3) immediately imply (2.2).

Finally, to establish (2.4), we note that by (A.3) with the appropriate substitutions we have for each i, k with $i \in S_k$, that $p_i a_{ik} = E(b_{sk}) p'_{ik}$, which we combine with (2.21).

Proof of (2.23) and (2.24). We note that $p_j = m / M$ for all $j \in S$; $p_j / p_j = p_i / p_i$ and $r_{j1} = r_{i1}$ for all j for which $A_j \in I_t$; the summation in the numerators of (2.8) (2.9) and (2.13) for $k = 1$ are over M_t, m_t and $\max\{M_t - N_t + n_t, 1\}$ elements, respectively; and that the summation in the denominator of these three equations is over $\min\{n_t, M_t\}$ elements. We then combine these relations with (2.8), (2.9), (2.13-2.16), (2.19-2.21) to obtain that $k' = 1$ (since $S_2 = \varnothing$) and that (2.23) and (2.24) hold.

REFERENCES

Brick, J. M., Morganstein, D. R., and Wolter, C. L. (1987), "Additional Uses for Keyfitz Selection," in *Proceedings of the Survey Research Section, American Statistical Association*, pp. 787-791.

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.

Ernst, L. R. (1986), "Maximizing the Overlap Between Surveys When Information Is Incomplete," *European Journal of Operational Research*, 27, 192-200.

Ernst, L. R. and Ikeda, M. (1994), "A Reduced-Size Transportation Algorithm for Maximizing the Overlap Between Surveys," Bureau of the Census, Statistical Research Division, Research Report Series, No. RR-93/02.

Gilliland, P. D. (1984), "1985 PATC Measures of Size Specifications," memorandum to W. Brown, Bureau of Labor Statistics, dated September 7.

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

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

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