SIMULTANEOUS SELECTION OF PRIMARY SAMPLING UNITS FOR TWO DESIGNS

Lawrence R. Ernst, Bureau of the Census*
Washington, DC 20233

KEY WORDS: Sampling, overlap, stratifications, maximizes

1. Introduction

Consider the following sampling problem. Primary sampling units (PSUs) are to be selected for two designs, denoted as designs 1 and 2, both of which are one PSU per stratum designs. The selection of sample PSUs 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 universe of PSUs is the same for both designs, but each is stratified independently. The sample PSUs in design 1 are required to be a subset of the sample PSUs in design 2. This necessitates the following assumption:

> The probability of selection for each PSU in design 1 does not exceed the probability of selection of that PSU in design 2. (1.1)

In this paper we demonstrate how the two-dimensional controlled selection procedure of Causey, Cox and Ernst (1985) can be used to satisfy all the conditions of this problem, that is,

> There is one sample PSU in each design 1 and design 2 stratum, selected with the required probabilities. (1.2)

> Each design 1 sample PSU is a design 2 sample PSU. (1.3)

A particular application of this procedure, to the proposed expansion of the Current Population Survey (CPS), which motivated this work, is presented in Section 6. Some readers may wish to read the beginning of that section before proceeding further, to obtain an understanding of this motivation.

Recently this author has become aware of a more general result by Pruhs (1989), who considers the same problem without the assumption (1.1), and consequently (1.3) is not true in general. Instead, using a graph theory approach, he presents an algorithm for which (1.2) is satisfied and the following additional condition holds:

> The expected value for the number of sample PSUs common to the two designs is maximized and the actual number in common for any sample is always greater than the expected value minus one. (1.4)

Thus, Pruhs views the problem as one of maximizing the number of sample PSUs common to the designs when the sample PSUs are chosen for the two designs simultaneously. Previously, Causey, Cox and Ernst (1985) and Ernst (1986) presented optimal linear programming procedures for maximizing the number of sample PSUs in common to two designs when the two sets of sample PSUs are chosen sequentially. In general, choosing the two samples simultaneously permits a larger expected overlap, but in many applications it is not possible to select the samples simultaneously, such as when the two designs are for the same periodic survey, but the second design is a redesign of the first design done at a later date.

It is shown here that the problem considered by Pruhs can also be solved by the controlled selection procedure of Causey, Cox and Ernst (1985). This approach has two advantages over Pruhs approach. The controlled selection approach involves a solution of a sequence of transportation problems. Commercial software is readily available which can solve transportation problems, 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. By contrast, both the theory and the task of programming the algorithm with Pruh's graph theory approach appears to be much more complex.

In Section 2, a brief review of the procedure of Causey, Cox and Ernst is given. In Section 3, the particular formulation of the sampling problem is presented. The presentation will first be for the more general problem in which (1.1) is not assumed. It will then be shown, quite simply, that with assumption (1.1), a special case of the general problem arises for which (1.3) is satisfied. In Section 4, the methods of avoiding some difficulties in using this procedure relating to rounding are described. In Section 5, formulas for the between PSU variance for linear estimates for both designs are presented for the controlled selection procedure. Finally, in Section 6, the application of the procedure to the proposed expansion of the CPS is considered, which includes an empirical comparison, for each design, of between PSU variances for controlled selection and independent selection.

Due to lack of space, some portions of the complete paper are omitted here. Specifically, Section 4, three of the four tables and the list of references have been omitted. The complete paper is available from the author.

2. Review of Controlled Rounding and Controlled Selection Concepts

The concepts of zero-restricted controlled rounding and controlled selection are briefly reviewed here. The reader is referred to Cox and Ernst (1982) and Causey, Cox and Ernst (1985) for more details and motivation on this subject, and

for other references.

An (m+1)×(n+1) array, $A=(a_{ij})$, is said to be a tabular array if

$$\sum_{i=1}^{m} a_{ij} = a_{(m+1)j}, \qquad j=1,...,n+1,$$

$$\sum_{j=1}^{n} a_{ij} = a_{i(n+1)}, \qquad i=1,...,m+1.$$

Such an array can be represented in the form

a_{11}	.	.	a_{1n}	$a_{1(n+1)}$
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
a_{m1}	.	.	a_{mn}	$a_{m(n+1)}$
$a_{(m+1)1}$	.	.	$a_{(m+1)n}$	$a_{(m+1)(n+1)}$

with the internal, row total, column total and grand total cells clear from this diagram.

A zero-restricted controlled rounding of an (m+1)×(n+1) tabular array, $A=(a_{ij})$, with respect to a positive integer base b is an (m+1)×(n+1) tabular array, $R(A) = (r_{ij})$, for which

$$r_{ij} = \lfloor a_{ij}/b \rfloor b \text{ or } \lceil a_{ij}/b \rceil b \text{ for all } i,j,$$

where $\lfloor x \rfloor$, $\lceil x \rceil$ denote the greatest integer not exceeding x and the smallest integer not less than x, respectively. If no base is stated, base 1 is understood.

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 tabular array.

If $S=(s_{ij})$ is an (m+1)×(n+1) tabular array, a solution to the controlled selection problem S is a finite sequence of arrays, $N_1 = (n_{ij1})$, $N_2 = (n_{ij2})$,..., $N_l = (n_{ijl})$, and associated probabilities, $p_1,...,p_l$, satisfying:

N_k is a zero-restricted controlled rounding of S for all k, (2.1)

$$\sum_{k=1}^{l} p_k = 1, \tag{2.2}$$

$$E(n_{ijk} \mid i,j) = \sum_{k=1}^{l} n_{ijk}\, p_k = s_{ij},$$

$$i=1,...,m+1,\ j=1,...,n+1. \tag{2.3}$$

If S arises from a sampling problem for which s_{ij} is the expected number of sampling units selected in each cell, and the actual number selected in each cell is determined by choosing one of the N_k's with its associated probability, then by (2.1) the deviation of s_{ij} from the number of sampling units actually selected from cell (i,j) is less than 1, whether i,j is an internal cell or a total cell. By (2.3) the expected number of sampling units selected is s_{ij}.

In Causey, Cox and Ernst (1985) a solution to the controlled selection problem is obtained by recursively defining the sequences $N_1,...,N_l$ and $p_1,...,p_l$ as follows. For fixed k, to define N_k, p_k, begin with the tabular array $A_k = (a_{ijk})$. $A_1 = S$ and for k≥1, A_{k+1} is defined in terms of N_k, p_k. N_k is simply a zero-restricted controlled rounding of A_k. To define p_k, first let

$$d_k = \max\{\, | n_{ijk} - a_{ijk} | : i=1,...,m+1,\ j=1,...,n+1\}, \tag{2.4}$$

and then let

$$p_k = 1 - d_k \quad \text{if } k=1$$

$$= (1 - \sum_{i=1}^{k-1} p_i)\,(1-d_k) \text{ if } k>1. \tag{2.5}$$

If $d_k>0$ define A_{k+1} by letting

$$a_{ij(k+1)} = n_{ijk} + (a_{ijk} - n_{ijk})/d_k \tag{2.6}$$

for all i,j, and then proceed to define N_{k+1}, p_{k+1}.

It is shown in Causey, Cox and Ernst (1985) that there is an integer l for which $d_l=0$ and that this terminates the algorithm; that is $N_1,...,N_l$ and $p_1,...,p_l$ satisfy (2.1)-(2.3).

3. The Controlled Selection Procedure for Selection of Sample PSUs

The procedure begins by construction of an (m+1)×(n+1) tabular array, S, for which a sequence of arrays, $N_1,...,N_l$, and associated probabilities, $p_1,...,p_l$, satisfying (2.1)-(2.3) lead to a solution of the problem described in the Introduction. To construct S, let m′, n′ denote the number of strata in designs 1 and 2, respectively, and let m=m′+1, n=n′+1. Let G_1, G_2 denote the random sets consisting of all sample PSUs in designs 1 and 2, respectively. For i=1,...,m′, j=1,...,n′, let t_{ij} denote the number of PSUs in the i-th design 1 stratum and j-th design 2 stratum; let B_{iju} denote the u-th such PSU, u=1,...,t_{ij}; and let T denote the set of all triples (i,j,u). For (i,j,u) ε T, let $P_{iju\alpha} = P(B_{iju} \in G_\alpha)$, α=1,2 and let $P_{iju3} = \min\{P_{iju1}, P_{iju2}\}$. Finally, for i=1,...m′, j=1,...,n′, let

$$s_{ij} = \sum_{u=1}^{t_{ij}} P_{iju3}, \tag{3.1}$$

$$s_{mj} = 1 - \sum_{i=1}^{m'} \sum_{u=1}^{t_{ij}} P_{iju3}, \tag{3.2}$$

$$s_{in} = 1 - \sum_{j=1}^{n'} \sum_{u=1}^{t_{ij}} P_{iju3}, \tag{3.3}$$

$$s_{mn} = 0, \tag{3.4}$$

and let $S=(s_{ij})$ denote the $(m+1)\times(n+1)$ tabular array with internal elements defined by (3.1)-(3.4). Note that the marginal values for S are as follows:

$$s_{i(n+1)} = 1, \quad i=1,...,m', \quad s_{(m+1)j} = 1, \quad j=1,...,n', \tag{3.5}$$

$$s_{m(n+1)} = n' - \sum_{(i,j,u)\in T} P_{iju3}, \tag{3.6}$$

$$s_{(m+1)n} = m' - \sum_{(i,j,u)\in T} P_{iju3}, \tag{3.7}$$

$$s_{(m+1)(n+1)} = m' + n' - \sum_{(i,j,u)\in T} P_{iju3}. \tag{3.8}$$

Interpretation of the array S will now be provided. For i=1,...,m', j=1,...,n', s_{ij} is the probability that a PSU is in sample for both designs, that is, in the i-th design 1 stratum and j-th design 2 stratum; while s_{mj} is the probability that the sample PSU in the j-th design 2 stratum is not a design 1 sample PSU, and s_{in} is the probability that the sample PSU in the i-th design 1 stratum is not a design 2 sample PSU. Thus, cells (i,j) for which $i\le m'$, $j\le n'$ can be thought of as corresponding to the selection of sample PSUs that are in both designs, while internal cells in row m correspond to the sample PSUs in design 2 only, and similarly, internal cells in column n correspond to design 1 only sample PSUs.

As for the marginals (3.5)-(3.8), (3.5), arises because there is one sample PSU in each design 1 and design 2 stratum. (3.6) indicates the expected number of PSUs to be selected as design 2 sample PSUs which are not design 1 sample PSUs, with an analogous interpretation for (3.7). (3.8) is the expected number of distinct PSUs that are to be in sample for at least one of the two designs.

After computing a set of arrays, N_k, and associated probabilities, p_k, k=1,...,*l*, satisfying (2.1)-(2.3) for this S using the controlled selection algorithm described in Section 2, the selection of the sample PSUs for the two designs is a two step process. First, one of the N_k's is selected. The internal cells of N_k are either 0 or 1. A 1 in cell (i,j) with $i\le m'$, $j\le n'$, indicates $B_{iju} \in G_1 \cap G_2$ for a single u=1,...,t_{ij}. Among the t_{ij} such PSUs, one is selected at the second step with conditional probability

$$P(B_{iju} \in G_1 \cap G_2 \mid n_{ijk}=1) = P_{iju3}/s_{ij}, \quad u=1,...,t_{ij}. \tag{3.9}$$

A 1 in cell (m,j), j=1,...,n', indicates that the sample PSU selected for design 2 from the j-th stratum is not to be a design 1 sample PSU. Among the $\sum_{i=1}^{m'} t_{ij}$ PSUs in the j-th design 2 stratum, one is selected at the second step with conditional probability

$$P(B_{iju} \in G_2 \sim G_1 \mid n_{mjk}=1) = (P_{iju2} - P_{iju3})/s_{mj}, \quad i=1,...,m', \quad u=1,...,t_{ij}. \tag{3.10}$$

An analogous expression holds for a 1 in an internal cell in column n.

This two-step procedure just described satisfies (1.2) and (1.4). To establish (1.2), first note that clearly, by (3.5), there is exactly one sample PSU in each design 1 and 2 stratum. To show that each PSU is selected into the design 1 and design 2 samples with the correct probabilities, observe that by (2.3), (3.9) and (3.10), it follows that for each $(i,j,u) \in T$,

$$P(B_{iju} \in G_1 \cap G_2) =$$

$$P(n_{ijk}=1)\, P(B_{iju} \in G_1 \cap G_2 \mid n_{ijk}=1) = P_{iju3},$$

$$P(B_{iju} \in G_2 \sim G_1) =$$

$$P(n_{mjk}=1)\, P(B_{iju} \in G_2 \sim G_1 \mid n_{mjk}=1) = P_{iju2} - P_{iju3}.$$

Consequently, $P(B_{iju} \in G_2) = P_{iju2}$. Similarly, it can be shown that $P(B_{iju} \in G_1) = P_{iju1}$. Hence, (1.2) holds.

To establish (1.4), first note that for any selection procedure satisfying (1.2),

$$P(B_{iju} \in G_1 \cap G_2) \le P_{iju3}, \quad (i,j,u) \in T,$$

and hence,

$$E[\text{card}\,(G_1 \cap G_2)] \le \sum_{(i,j,u)\in T} P_{iju3}.$$

Then (1.4) follows, since for the current procedure, (2.3) and (3.1) yield

$$E[\text{card}\,(G_1 \cap G_2)] = \sum_{i=1}^{m'} \sum_{j=1}^{n'} E(n_{ijk} \mid i,j) = \sum_{(i,j,u)\in T} P_{iju3},$$

and (2.1), (3.5), (3.6) yield

$$\text{card}\,(G_1 \cap G_2 \mid N_k) = \sum_{i=1}^{m'} \sum_{j=1}^{n'} n_{ijk} = n' - n_{m(n+1)}$$

$$> \sum_{(i,j,u)\in T} P_{iju3} - 1, \quad k=1,...,l.$$

Finally, to show (1.3) holds for this procedure with the additional assumption (1.1), simply observe that if $P_{iju1} \le P_{iju2}$ for all $(i,j,u)\in T$, then by (3.3),

$$s_{in} = 1 - \sum_{j=1}^{n'} \sum_{u=1}^{t_{ij}} P_{iju1} = 0, \quad i=1,...,m',$$

and hence $G_1 \sim G_2 = \emptyset$ for all samples. Note that in this

case, the n-th column can be omitted in defining S.

5. Variances for the Controlled Selection Procedure

In this section variance formulas are derived for estimators of total for both designs when using the sample procedure detailed in Section 3, under the assumption that a census is conducted in the sample PSUs. If the sample PSUs are subsampled then these formulas represent the between PSU component of variance. Let X denote the total value over the entire population for a characteristic of interest, and let X_{iju} denote the total for PSU B_{iju} for each (i,j,u) ε T. For α=1,2, let $\hat{X}_\alpha$ denote the usual estimator for X for design α corresponding to probability proportional to size sampling, that is

$$\hat{X}_\alpha = \sum \frac{X_{iju}}{P_{iju\alpha}},$$

where the summation is over all (i,j,u) such that $B_{iju} \varepsilon G_\alpha$.

For (i,j,u), $(i^*,j^*,u^*) \varepsilon$ T, $(i,j,u) \neq (i^*,j^*,u^*)$, α = 1,2, let

$$\pi_{ijui^*j^*u^*\alpha} = P(B_{iju}, B_{i^*j^*u^*} \varepsilon G_\alpha).$$

Then from Raj (1968, p.54),

$$V(\hat{X}_\alpha) = \frac{1}{2} \times$$

$$\sum_{\substack{(i,j,u),(i_*,j_*,u_*)\varepsilon T \\ (i,j,u)\neq(i^*,j^*,u^*)}} (P_{iju\alpha} P_{i^*j^*u^*\alpha} - \pi_{ijui^*j^*u^*\alpha}) \left(\frac{X_{iju}}{P_{iju\alpha}} - \frac{X_{i^*j^*u^*}}{P_{i^*j^*u^*\alpha}}\right)^2 \qquad (5.1)$$

Consequently, it is only necessary to show how to compute $\pi_{ijui^*j^*u^*\alpha}$ for each (i,j,u), $(i^*,j^*,u^*) \varepsilon$ T, $(i,j,u)\neq(i^*,j^*,u^*)$. To do this for α=2, first observe that $\pi_{ijui^*j^*u^*2}$=0 if $j=j^*$. Consequently, it may be assumed from now on that $j\neq j^*$. For each such i,j,i^*,j^*, let $r_{iji^*j^*} = P(n_{ijk} = n_{i^*j^*k} = 1)$. Note that $r_{iji^*j^*}$ is the sum of p_k over all k for which $n_{ijk} = n_{i^*j^*k} = 1$. Then to obtain $\pi_{ijui^*j^*u^*2}$, observe that both B_{iju} and $B_{i^*j^*u^*}$ can be in G_2 if either

$n_{ijk} = n_{i^*j^*k} = 1$, $n_{mjk} = n_{i^*j^*k} = 1$, $n_{ijk} = n_{mj^*k} = 1$ or

$n_{mj} = n_{mj^*k} = 1$,

which combined with (3.9) and (3.10) yield the four terms in the following expression:

$$\pi_{ijui^*j^*u^*2} = r_{iji^*j^*} \frac{P_{iju3}}{s_{ij}} \frac{P_{i^*j^*u^*3}}{s_{i^*j^*}}$$

$$+ r_{mji^*j^*} \frac{(P_{iju2} - P_{iju3})}{s_{mj}} \frac{P_{i^*j^*u^*3}}{s_{i^*j^*}}$$

$$+ r_{ijmj^*} \frac{P_{iju3}}{s_{ij}} \frac{(P_{i^*j^*u^*2} - P_{i^*j^*u^*3})}{s_{mj^*}}$$

$$+ r_{mjmj^*} \frac{(P_{iju2} - P_{iju3})}{s_{mj}} \frac{(P_{i^*j^*u^*2} - P_{i^*j^*u^*3})}{s_{mj^*}}. \qquad (5.2)$$

The only differences in the expression for $\pi_{ijui^*j^*u^*1}$, which is obtained similarly, are that the subscripts mj, mj^*, and 2 are replaced by the subscripts in, i^*n, and 1, respectively, and that $\pi_{ijui^*j^*u^*1}$=0 if $i=i^*$.

Note, in the special case when $P_{iju1} \leq P_{iju2}$ for all (i,j,u) ε T, it follows that since $P_{iju3} = P_{iju1}$, then the last three terms in the expression for $\pi_{ijui^*j^*u1}$ drop out, and hence

$$\pi_{ijui^*j^*u^*1} = r_{iji^*j^*} \frac{P_{iju1}}{s_{ij}} \frac{P_{i^*j^*u^*1}}{s_{i^*j^*}}.$$

All four terms in $\pi_{ijui^*j^*u2}$ remain, although now 1 can be substituted for 3 in (5.2).

Note that (5.2), and hence (5.1), are different for the controlled selection procedure than for independent sampling for each design. In the latter case, $\pi_{ijui^*j^*u^*\alpha} = P_{iju\alpha} P_{i^*j^*u^*\alpha}$ if either α=1 and $i\neq i^*$, or if α=2 and $j\neq j^*$, and hence there is no between stratum component of variance for independent sampling. An empirical comparison of the variances for the two procedures for one application is presented in the next section.

6. APPLICATION TO PROPOSED EXPANSION OF THE CURRENT POPULATION SURVEY

A potential important application of the controlled selection procedure described in the preceding sections is to the proposed "two-phase" expansion of the Current Population Survey (CPS). 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, based on 1990 census data, is scheduled to be phased in. The reliability requirements for the redesign are expected to be approximately the same as in the present design. Beginning in 1996, if the proposal is implemented in its present form, a sample expansion will take place to meet strengthened reliability requirements, which will enable monthly estimates to be released for all 50 states and the District of Columbia. Presently annual estimates for all states are released and monthly estimates only for the 11 largest states, in addition to monthly national estimates.

Each month the expanded sample will be interviewed over the two-week period consisting of the weeks containing the 19th and 26th of the month, instead of only the single week containing the 19th, as at present. A portion of the total sample will be designated as the national sample. This sample will be interviewed during the first interview week and will be used in the national estimates. The remaining sample, designated as the state supplement sample, will be interviewed mainly during the second week. The national sample and the state supplement sample will be used together to produce the monthly estimates for all states. The expanded sample has been designated as the "two-phase" sample because it consists of both a national and a state supplement sample. The redesign prior to the expansion has been denoted D_{90} (for 90's redesign) at the Census Bureau, while the expanded design has been denoted D_2 (for two-phase design).

The drawbacks associated with most approaches to the selection of sample PSUs for these two designs present a key problem in attempting to obtain optimal sample designs for both D_{90} and D_2. For example, if the D_{90} sample PSUs are selected first from an optimal D_{90} stratification and then additional sample PSUs are selected to join the D_{90} sample PSUs to form the set of D_2 sample PSUs, variances for the D_2 design will generally be higher than if the D_2 sample PSUs are selected directly from an optimal D_2 stratification. (As in Section 5, "variances" in this section refers only to the between PSU component.) Similarly, a suboptimal procedure for D_{90} PSU selection will result if the D_{90} sample PSUs are obtained by subsampling D_2 sample PSUs selected from an optimal D_2 stratification. Finally, although increases in variances for either design can be avoided by independently selecting D_{90} and D_2 sample PSUs from optimal D_{90} and D_2 stratifications, this approach will generally result in some D_{90} sample PSUs being dropped from the D_2 sample, a feature which undesirably impacts on field operations. Some of these approaches are discussed in Chandhok, Weinstein and Gunlicks (1990).

The controlled selection approach of this paper with assumption (1.1) can be used as a procedure for simultaneously selecting sample PSUs for both designs while avoiding all of these problems. To use this procedure, first obtain optimal stratifications for D_{90} and D_2, which correspond to the design 1 and design 2 stratifications respectively in the terminology used in the previous sections. Then the controlled selection procedure results in a set of sample PSUs for D_{90} and D_2 satisfying (1.2) and (1.3).

As noted in Section 5, the variances for estimates obtained for the D_{90} and D_2 designs with controlled selection differ from the variances that are obtained if the sample PSUs are selected independently for each design. An empirical investigation was undertaken to compare variances using these two approaches to PSU selection.

For the comparison of the variances, the D_{90} and D_2 stratifications were obtained using several labor force characteristics from the 1980 census as stratification variables. 1980 census data were substituted for the yet unavailable 1990 data. A modified Freedman-Rubin clustering algorithm (Kostanich et al. 1981) was used to obtain the stratifications. The D_{90} and D_2 stratifications and the controlled selection were performed separately for each state, since it has been shown that for each design a sample meeting reliability requirements for each state would also meet the reliability requirements for national estimates.

The variables used here to compare the independent selection and controlled selection variances are number of unemployed persons and number of persons in the civilian labor force. The comparisons were done only for the 31 states listed in Tables 1-4. Of the remaining 20 states (counting the District of Columbia), the 11 largest were omitted since the precision requirements for this study, and hence the stratifications, were the same for D_{90} and D_2. Eight states were omitted because they consisted entirely of self-representing PSUs for D_2. For these 19 states, variances for controlled selection and independent selection would be identical for both D_{90} and D_2. Finally, Alaska was omitted because of problems with the data files.

For each state and each characteristic, variances were computed for each of the two designs and each of the two selection procedures, using both 1980 and 1970 census data. The 1980 data were used to compare variances for variables at the same point in time for which the stratification was done, while the 1970 data were used to simulate a 10-year lag between the data used in the stratifications and the collection of the survey data, which is roughly the anticipated average lag time for the D_{90} and D_2 designs.

The variances for the controlled selection procedure for D_{90} and D_2 with 1980 data are presented in Table 1 for 1980 data and in Table 2 for 1970 data. Both tables are omitted here.

Tables 3 and 4 can be used to compare the variances for the controlled selection and independent selection procedures for 1980 and 1970 data, respectively. Table 3 is omitted here. For each state, the ratio of the variance for the controlled selection procedure to the variance for independent selection for the indicated characteristic and design is presented in numerical columns 1,2,4 and 5. Each entry in the next-to-last row of the column is the arithmetic mean of the entries in the preceding rows of that column. Each entry in the last row is the ratio of the variance for controlled selection to the variance for independent selection for the total number of persons in all the listed states with the indicated characteristic. Finally, each entry in columns 3 and 6 is the arithmetic mean of the entries in the preceding two columns of that row.

The deviations of the ratios of the variances from 1 are generally numerically smaller for D_2 than for D_{90} on a state-by-state basis, and the deviation from 1 for D_2 for all states combined, (in the last two rows of the tables) is quite small. As for D_{90}, although the ratios in the majority of the columns are less than 1 for at least half the states, the bottom two rows of these tables are not particularly favorable to the controlled selection procedure, particularly the final row for 1980 data, whose entries are numerically larger than the row above it due to the presence of a few states with large variances together with large ratios of the variances, a combination which increases the last row much more than the next to last row. Elaboration on these observations will be given in the remainder of this section. (Note that since

the characteristics that appear in these tables were asked in the censuses only for a sample of the population, the "variances" used in computing the ratios in the tables are only estimates of the between PSU variances. Consequently, any comparisons made are not statistical inferences applying to an entire census universe.)

There are at least two explanations for the smaller deviations from 1 for D_2 of the ratios of the variances. First, for the controlled selection procedure, no two PSUs in the same D_2 stratum can be sample PSUs for the D_{90} design, while there can be as many D_2 sample PSUs from a single D_{90} stratum as there are D_2 strata containing PSUs from that D_{90} stratum. Thus the restrictions imposed by the controlled selection procedure on the possible sets of sample PSUs are more restrictive for D_{90} than for D_2, which partially explains the smaller deviations of the ratios for D_2.

The second reason for the smaller deviations for D_2 is that many of the D_2 strata consist entirely of PSUs from a single D_{90} stratum. If the j-th D_2 stratum is such a stratum, then $\pi_{ijui^*j^*u^*2} = P_{iju2}\, P_{i^*j^*u^*2}$ if $j^* \neq j$ and $\pi_{ijui^*j^*u^*2} = 0$ if $j^*=j$, for all distinct pairs of triples with j fixed, (i,j,u), $(i^*,j^*,u^*) \varepsilon T$, for both controlled selection and independent selection, and thus the contribution to (5.1) from all such pairs is the same for both procedures for D_2. No analogous relationship holds for D_{90}.

We now consider further the question of whether controlled selection or independent selection should yield lower variances. Note that for each $(i,j,u) \varepsilon T$,

$$\sum_{(i^*,j^*,u^*)\varepsilon T,\ (i^*,j^*,u^*)\neq(i,j,u)} \pi_{ijui^*j^*u^*\alpha} = (m'-1)\, P_{iju1} \quad \text{if } \alpha=1,$$

$$= (n'-1)\, P_{iju2} \quad \text{if } \alpha=2,$$

for both controlled selection and independent selection (see Raj (1968)), and hence

$$\sum_{(i,j,u),(i^*,j^*,u^*)\varepsilon T,\ (i,j,u)\neq(i^*,j^*,u^*)} (P_{iju\alpha}\, P_{i^*j^*u\alpha} - \pi_{ijui^*j^*u^*\alpha})$$

are the same for both procedures. Consequently, there is no reason to expect the variances for one procedure to be higher or lower than the other unless the relationship between the $P_{iju\alpha}\, P_{i^*j^*u^*\alpha} - \pi_{ijui^*j^*u^*\alpha}$ and the $(X_{iju}/P_{iju\alpha} - X_{i^*j^*u^*}/P_{i^*j^*u^*\alpha})^2$ factors differs for the two procedures. Actually, it was surmised prior to performing the computations that controlled selection might yield lower variances than independent selection for D_{90}. This is because $\pi_{ijui^*j^*u^*1} = 0$ for controlled selection if $j^*=j$, while $(X_{iju}/P_{iju\alpha} - X_{i^*j^*u^*}/P_{i^*j^*u^*\alpha})^2$, $\alpha=1,2$, tends to be small for such pairs of PSUs since they are both in the same D_2 stratum. Thus, controlled selection may result in many pairs of PSUs with a large value for $P_{iju1}\, P_{i^*j^*u^*1} - \pi_{ijui^*j^*u^*1}$ and a small value for $(X_{iju}/P_{iju1} - X_{i^*j^*u^*}/P_{i^*j^*u^*1})^2$, a combination which tends to lower variances. The data in the tables fail to support this supposition, however.

In summary, controlled selection retains all D_{90} sample PSUs for the D_2 design and selects the sample PSUs for both designs from their optimal stratifications. Controlled selection appears at least for this limited study, to yield variances quite close to independent selection for D_2. More study may be needed on the effects of controlled selection on the variances for D_{90}. If the results are favorable for controlled selection, it appears to be a contender for the PSU selection procedure for the proposed CPS expansion. Note, however, that unless one is willing to ignore the effect on the variances of the between stratum variance component induced by controlled selection, variance estimation will be more complex than for some other approaches to PSU selection.

Acknowledgement

The author would like to thank Todd Williams who did the programming that produced all the tables and figures that appear in this paper.

* This paper reports the general results of research undertaken by Census Bureau staff. The views expressed are attributable to the authors and do not necessarily reflect those of the Census Bureau.

Table 4. Ratios of Controlled Selection Variances to Independent Selection Variances for 1970 Data

State	D_{90}			D_2		
	Total Unemployed	Total in Civilian Labor Force	Mean of Ratios of Variances	Total Unemployed	Total in Civilian Labor Force	Mean of Ratios of Variances
Alabama	0.711	1.237	0.974	1.122	1.032	1.077
Arizona	0.828	0.614	0.721	1.074	0.885	0.979
Arkansas	0.878	1.436	1.157	1.061	0.987	1.024
Colorado	1.601	0.783	1.192	1.014	0.997	1.006
Georgia	0.800	1.198	0.999	1.041	1.016	1.028
Idaho	0.665	0.875	0.770	1.099	1.063	1.081
Indiana	0.567	1.338	0.953	0.892	1.472	1.182
Iowa	1.226	1.602	1.414	0.783	0.829	0.806
Kansas	0.728	0.986	0.857	0.955	0.923	0.939
Kentucky	2.075	0.418	1.246	1.083	0.838	0.960
Louisiana	2.273	0.764	1.518	0.949	0.972	0.960
Maryland	0.913	0.851	0.882	1.000	1.000	1.000
Minnesota	0.884	0.819	0.851	1.117	0.875	0.996
Mississippi	1.116	1.661	1.389	0.933	1.028	0.981
Missouri	0.555	0.553	0.554	0.950	1.089	1.020
Montana	0.822	1.235	1.029	0.881	1.188	1.034
Nebraska	0.946	0.776	0.861	0.991	1.001	0.996
Nevada	1.118	0.655	0.886	1.023	0.862	0.943
New Mexico	0.747	0.628	0.687	0.908	1.405	1.156
North Dakota	0.712	0.882	0.797	0.908	0.723	0.815
Oklahoma	1.141	0.430	0.786	1.019	0.834	0.927
Oregon	0.737	0.632	0.684	0.894	0.975	0.934
South Carolina	1.252	0.834	1.043	1.166	1.099	1.133
South Dakota	0.853	0.993	0.923	0.898	0.951	0.925
Tennesse	1.105	0.578	0.842	1.097	1.002	1.050
Utah	0.933	1.479	1.206	0.966	0.936	0.951
Virginia	1.106	2.103	1.604	0.952	1.330	1.141
Washington	1.144	0.536	0.840	0.937	1.247	1.092
West Virginia	3.049	2.130	2.590	0.932	1.125	1.029
Wisconsin	1.604	0.919	1.262	0.958	0.980	0.969
Wyoming	0.365	0.433	0.399	0.671	1.029	0.850
Mean of Ratios of Variances	1.079	0.980	1.030	0.977	1.022	0.999
Ratios of Variances for Sum of State Totals	1.135	1.026	1.080	0.990	1.084	1.037