THE EFFECT OF CONTROLLED SELECTION ON THE BETWEEN PSU VARIANCE

Camilla Brooks, Bureau of the Census

Introduction

In the 1970 Current Population Survey (CPS) redesign the primary sampling units were selected by the method of controlled selection, rather than the independent selection of a PSU from each stratum. The expectation has been that the control of the geographic dispersion of the sample of primary sampling units by states (the primary control used) would produce a lower contribution to variance for many CPS items than would result from an independent selection. This paper gives the results to date of a study of some characteristics of a regional, U.S. and state nature. A brief explanation of both the CPS design and the controlled selection procedure is also presented.

The Current Population Survey (CPS)

The CPS is a national survey conducted monthly by the Bureau of the Census to provide national estimates of labor force characteristics as well as other demographic characteristics.

This monthly survey which covers the civilian noninstitutional population of the U.S. consists of about 47,000 eligible households in 461 primary sampling units (PSU's) selected from 376 strata. (A PSU consists of one or more contiguous counties.). Of these 376 strata 156 of them are considered self-representing; i.e., they consist of only one PSU. The other 305 PSU's are considered nonself-representing (NSR) and are selected from the 220 remaining strata.

The 461 PSU Current Population Survey design is a combination of two independent samples, referred to as the A sample and C sample. The 156 self-representing PSU's are part of both the A and C sample designs. The NSR portion of the A sample consists of one PSU selected from each of the 220 NSR strata; the NSR portion of the C sample consists of the selection, independently of the A sample selection, of one PSU from each of 110 paired A sample strata. Thus the A sample design of the CPS consists of 376 PSU's, the C sample design of 266 PSU's, with 25 NSR PSU's in both the A and C sample designs. This study concerns itself with the nonself-representing portion of the A sample design only.

In practice, the CPS is a multi-stage survey; that is, given the sample of PSU's, a sample of segments (consisting of groups of households) is selected from each sample PSU. The present study ignores the multi-stage nature of the CPS and assumes complete coverage of the sample PSU's. Thus the difference in variance of the estimate of total is due only to the difference in the method of selecting the NSR PSU's.

Controlled Selection

Controlled selection is a technique used in probability sampling to increase the probability of selecting samples consisting of certain preferred combinations of sampling units, while maintaining pre-specified probabilities of selection for individual units in the population. Consequently, the probabilities of less desirable combinations are reduced, sometimes to zero [2]. Though it is hoped that its use will reduce the variance of at least a number of items, it is not always used expressly for this purpose. Such was the case in the selection of PSU's for the CPS. Though it was expected that the use of controlled selection would reduce the between PSU contribution to variance, the primary reasons for its application were political and budgetary.

The conditions imposed on the designation of acceptable patterns in the controlled selection procedure are referred to as controls. An acceptable pattern consists of a set of sampling units, one selected from each stratum, such that the prespecified controls are met. There is considerable subjectivity in the selection of each acceptable pattern. However, it is a probability selection rather than a purposive selection when a complete set of patterns is designated and a single pattern is selected with the probability assigned to the pattern. Every sampling unit is in at least one pattern and the probabilities of the patterns in which a particular unit appears must add to its assigned probability of selection. A complete set of patterns such that the probabilities accumulate to 1.0000 is only a subset of all possible patterns [1].

Every 10 years the CPS undergoes major design revisions incorporating information from the most recent decennial census. PSU's (which may or may not have been changed) are restratified and one PSU is selected from each stratum. In the most recent redesign the nonself-representing PSU's were selected using a Goodman-Kish controlled selection procedure [1]. Two controls were instituted: (1) the number of PSU's selected from each state group and (2) the number of PSU's in the then current sample to be retained--these referred to as "overlapping PSU's". The first control was used to ensure that each state was represented with a number of nonself-representing PSU's proportional to the NSR population in the state. To do this the probabilities of all PSU's within a state were summed over all the strata within a region. This sum gave the expected number of PSU's to be selected from the state. (It is herein referred to as the expected value.). The second control was instituted to ensure the selection of the number of 1960 design PSU's appropriate to the expected value of the "overlapping PSU's." The probabilities assigned to the PSU's in 1970 were computed using an unbiased method developed by Keyfitz [3] and extended by W. M. Perkins [4] and [5]. This procedure maximizes the retention of the old design PSU's by taking into account the 1960 sample PSU's and their probabilities of selection. This was desirable mainly for budgetary reasons.

The controls used for this study were derived in the same manner as those used in the 1970 CPS sample; however, the probabilities assigned to each PSU represent their proportion of the stratum population rather than the Keyfitz probabilities. The experiment therefore does not reflect the impact of maximizing the overlap of 1960 and 1970 sample PSU's. These two factors, the state groups and the overlap, are referred to as the control groups; their expected values determine the number of PSU's to be selected from each control group in each acceptable pattern. As an example, NYA--control group 5--had an expected value of 2.8181; therefore, each pattern had to contain two or three PSU groups from this portion of New York State (two with probability of approximately .2 and three with probability of .8). Control 10--overlap--had an expected value 6.9121; thus either six or seven PSU's that were in sample in 1960 were assigned to each pattern.

Within each stratum the PSU's were clustered according to control considerations; i.e., all PSU's within a stratum with the same control designations, both state group and overlap/nonoverlap, were combined and treated as one PSU in the controlled selection program.

After a pattern was selected, the lowest probability of the PSU's selected or of the control group was assigned to the pattern. This probability was subtracted from every PSU group selected in the pattern and every control used (e.g., if New York A were selected twice, then that pattern probability was subtracted from the probability of selecting two PSU groups from this control). In other words, each time a pattern contained a PSU group or a control was used, a portion of the PSU group probability was absorbed. After identifying a complete set of controlled selection patterns, the sum of the probabilities of each pattern in which a PSU group was contained should have been equal to the original probability assigned the PSU group. In addition the sum of the probabilities of all the selected patterns would add to 1.0000.

In the 1970 design a specific pattern was selected with probability proportionate to its assigned probablity. Within each of the PSU groups in the selected pattern a PSU was selected with probability proportionate to size.

The controlled selection of PSU's was carried through independently in each region. The Northeast region, consisting of 20 strata was completed in 40 patterns; the West with 26 strata was completed in 108 patterns; the North Central with 72 strata was completed in 227 patterns; and the South with 102 strata was completed in 272 patterns.

Variance Estimation

Data used in this study were from the 1970 basic census record tapes with 20 percent sample data for the labor force items and 15 percent for the school enrollment items. For purposes of this study the PSU totals are assumed to be free of error. Persons in the Armed Forces were excluded.

To make the comparison between the controlled selection procedure and an independent selection procedure, the variances of the two procedures were examined. The controlled selection procedure affects only the between PSU part of the variance. The effect of this procedure on the total variability of a statistic depends on the relative sizes of the between and within PSU variances. For example, if the between PSU variance is very small in relation to the within PSU variance then even if controlled selection produces a large reduction in this variance component, it will not be meaningful.

Let

- X_{hi} = census total for the i-th PSU in the h-th stratum
- X_h = $\sum_{i}^{N_h} X_{hi}$ - the census total for the h-th stratum where N_h = number of PSU's in the h-th stratum
- P_{hi} = population in the i-th PSU in the h-th stratum
- P_h = population in the h-th stratum
- L = total number of strata in the region.

The between PSU variance for an estimate based on an independent selection of sample PSU's in each stratum is given by

$$\sigma^2_{IS} = \sum_{h}^{L} \sum_{i}^{N_h} \frac{P_{hi}}{P_h} \left[\frac{X_{hi}}{P_{hi}} P_h - X_h \right]^2 . \qquad (1)$$

To estimate the variance of the controlled selection procedure, a Monte Carlo procedure was used. For each replication a PSU was selected from the sample PSU group with probability equal to the proportion the PSU population was of the group population. The between PSU variance for the controlled selection of sample PSU's is given below:

$$\sigma^2_{CS} = \sum_{j=1}^{R} \left[\sum_{h}^{L} \sum_{i=1}^{1} \frac{X_{hij}}{P_{hij}} P_h - X \right]^2 \frac{1}{R} \qquad (2)$$

where

- R = number of replications
- X_{hij} = the census count of the i-th PSU (selected randomly with PPS from the PSU group) selected in the j-th replication
- X = census total for the region
- P_h = population of the h-th stratum
- P_{hi} = population of the i-th PSU in the h-th stratum.

The between PSU contribution to variance of the independent selection for state estimates is given by

$$\sigma^2_{IS(states)} = \sum_{h}^{L} \sum_{i}^{N_h} \frac{P_{hi}}{P_h} \left[\frac{X_{hi} \delta_{si}}{P_{hi}} P_h - X_{hs} \right]^2 \qquad (3)$$

where

δ_{si} = 1 if the i-th PSU in the h-th stratum is in the s-th state; 0 otherwise

X_{hs} = $\sum_{i}^{N_h} X_{hi}\,\delta_{si}$ - the census total of the s-th state in the h-th stratum where N_h = number of PSU's in the h-th stratum

The between PSU contribution to variance of state estimates when the controlled selection procedure is used is given by

$$\sigma^2_{CS(states)} = \sum_{j=1}^{R} \left[\sum_{h}^{L} \sum_{i=1}^{1} \frac{X_{hij}\delta_{si}}{P_{hij}} P_h - X_s\right]^2 \frac{1}{R}$$

where

X_s = census total for the s-th state in the region.

The efficiency of selecting the PSU's by the controlled selection procedure rather than the independent method is measured by the ratio

$$\sigma^2_{CS}/\sigma^2_{IS} .$$

In estimating the variance for the controlled selection of sample PSU's, 500 replications were used in the Northeast and West; 1,000 replications were used in the South and North Central, the two larger regions. Though it is felt that these number of replications provide a fairly stable variance estimate for the set of controlled selection patterns used, it does add unnecessary noise to the estimate. Subsequent to the use of this method, it was suggested that the following exact method of variance calculation be used. The variance is calculated as the sum of two terms--the variance attributable to the controlled selection of each PSU group and, given a particular pattern is selected, the variance among the PSU's in each selected PSU group. In the near future this method will be used for at least one region and the results compared to those of the procedure presented in this paper.

Discussion of Results

Table 1 gives the results of the study for the four regions and the total U.S. In obtaining the variances at the U.S. level it was simply assumed that the regions were independent. Variance ratios of the states in the West region are presented in Table 2. The variances of the controlled selection procedure are subject to error since they were estimated by a Monte Carlo technique. Since the analysis is based primarily on the observed variance ratios, it should be kept in mind that the observations are made without benefit of the knowledge of these errors. As used in the tables and in the text, data for blacks include that for races other than white whether or not specifically stated in the comparisons.

Some differences in regions are expected because of sampling variability, i.e., the variance resulting from the use of one set of patterns out of many possible sets of patterns. In addition differences result from a variable number of strata and PSU's in each region and a consequential variation in the number of controlled selection patterns in which each region is completed. Most of the variance ratios observed for the Northeast regional estimates are near or greater than 1.000. With a set of only 40 controlled selection patterns, it is expected that the controlled selection variance for this region would be subject to more error than the other regions. Also with only 20 NSR strata and 89 PSU's included in these strata, controlled selection probably would not have as much effect as in the other regions. The small population could also add to the differences.

The West, the next smallest region, had ratios comparable to the other regions. It is noted, however, that though the West has only 26 NSR strata there are 367 NSR PSU's so there is more "room for improvement" in the West than in the Northeast.

In regard to the states it is to be expected that a greater variability is attached to the variance ratios than would be in the regions. Nonetheless, many of the ratios are substantially under 1.000 and a considerable difference in variance between the two methods of PSU selection is indicated. This is expected since the primary control involved a proportionate selection of PSU's from each state.

States represented in a large proportion of the strata in the region would in general be expected to benefit more from controlled selection than those confined to a few strata. Alaska and Hawaii both are confined to one stratum with Alaska constituting an entire stratum. Thus the use of controlled selection should have no effect (and the ratio of the variances should be in the vicinity of one for all items). Variations from 1.000 should be within the error introduced by the Monte Carlo estimation. Nevada generally has somewhat higher variance ratios that the other states; it is not as widely dispersed over the region as some other states, e.g. New Mexico, Montana, and California.

The item population 14+ in itself is of little importance in discussing variance since it is subject to little relative sampling error. However, it is of some significance here in that the between PSU variance ratio for the regions (excepting the Northeast) and U.S. is smallest for this item. For the states, though it is not always the lowest ratio, it is one of the lowest ratios; the ratio of the two variances for population 14+ was .023 for New Mexico.

For the U.S. and regional estimates of school enrollment the ratios are in the vicinity of 1.000 with a somewhat lower ratio for blacks than whites. However, the individual states of the West display the opposite behavior. The ratio of the between PSU variance of the controlled selection procedure to the independent procedure is generally lower for whites than blacks.

Arizona, for example, has a ratio of .364 for whites, while that for blacks is 1.079.

For the U.S., West, and Northeast regions the variance ratios of the whites in the civilian labor force 14+ and employed 16+ are lower than the unemployment ratios; for the U.S. these ratios are about .65. Generally, the ratios of labor force items for blacks are higher than those for whites and those for females higher than those for males at the U.S. level; however the between PSU variance ratios for black females in the civilian labor force and unemployed are lower or about the same as those of black males. The South is the only region which shows an appreciable reduction in variance for whites employed in agriculture.

Thus for the U.S. it appears that the between PSU variance of estimates that are a greater portion of the population are reduced more by the use of the controlled selection procedure. For the two regions and many states, the ratios of unemployment items are as low as civilian labor force and employment items. For the states the variance ratios are relatively low for all total and white items. There is only a small percentage of blacks living in the West which could explain why many of the ratios for this population group are high.

How greatly the variances are affected by controlled selection will, of course, depend also on what proportion the between PSU variance is of the total variance. Many of the variance reductions for the total U.S. may be meaningless when considering the total variances. As an example of the reduction in total variance with the use of controlled selection the item "total unemployed" may be considered. It is estimated that the between PSU contribution to variance for this item is approximately 0.015. From Table 1 the ratio of the between PSU variance of controlled selection procedure to that of the independent selection is estimated as 0.818; therefore, the reduction in variance with the use of controlled selection would be approximately .015 - (.015x.818) = .003 or .3 percent of the total variance. Other items might produce a greater reduction in variance as a result of the use of controlled selection, but as is evident, for some the reduction will be trivial.

The impact of controlled selection on state estimates is likely to be much more important for two reasons. First, the proportion of the total variance for a state estimate contributed by sampling of PSU's is much larger--recall that CPS NSR strata typically include PSU's from more than one state. Second, the observed variance ratios indicate controlled selection has a much greater impact on the between PSU variance component, as evidenced by the lower variance ratios for the states. Looking at one of the states, New Mexico, for example, the ratio of the variances for total unemployment is .079. The reduction assuming the between PSU variance is one-third the total (this is reasonable for at least some of the states) would be approximately 30 percent. It should be noted that the proportion of the variance of the state estimate contributed by sampling of PSU's is subject to some control by modifying the estimation procedure. Thus for some items in some states the controlled selection procedure could affect the variance substantially. More knowledge of the between PSU contribution to total variance for states is necessary to judge its effect.

Comparisons of the between PSU contribution to variance for the two procedures of selecting PSU's, controlled selection and independent, will be done for states in other regions in the near future. As these are completed, greater insight into the effect of controlled selection on the between PSU variance will be gained.

References

[1] Goodman, R. and Kish, L., "Controlled Selection-A Technique in Probability Sampling", JASA, Volume 45, pp. 350-372, (1950).

[2] Jabine, Thomas B., "Controlled Selection - Statement of Problems", Draft Memorandum, Bureau of the Census, May 25, 1966.

[3] Keyfitz, Nathan, "Sampling with Probability Proportional to Size Adjustment for Changes in the Probabilities", JASA, (1951) Volume 46, pp. 105-109.

[4] Perkins, W.M., "1970 CPS Redesign: Proposed Method for Deriving Sample PSU Selection Probabilities Within 1970 NSR Strata", Memorandum to J. Waksberg, August 5, 1970.

[5] Perkins, W.M., "1970 CPS Redesign: Proposed Method for Deriving Sample PSU Selection Probabilities Within 1970 NSR Strata- Justification of Procedure as Unbiased", Memorandum to J. Waksberg, February 19, 1971.

[6] U.S. Bureau of the Census, The Current Population Survey - A Report on Methodology. Technical Paper Number 7, U.S. Government Printing Office, Washington, D.C., 1963.

TABLE 1. BETWEEN PSU VARIANCE RATIOS--CONTROLLED SELECTION TO INDEPENDENT SELECTION

(U.S. and Regions)

Characteristics	Total U.S. (1)	Northeast (2)	North Central (3)	South (4)	West (5)
NSR Population 14+	.340	1.264	.588	.485	.203
Enrolled in School 14-34					
Whites	1.149	1.531	1.292	.999	.998
Blacks and Others	.798	1.037	1.101	.744	.896
Males	1.092	1.488	1.232	.960	.921
Females	1.178	1.506	1.444	.947	.910
Labor Force Items					
In civilian labor force 14+					
Males, white	.675	1.010	.757	.690	.602
Males, black and other	.840	1.081	.916	.783	.961
Females, white	.699	.991	.755	.720	.540
Females, black and other	.779	1.004	.928	.722	.941
Whites	.645	.999	.713	.674	.537
Blacks and others	.808	1.041	.919	.748	.954
Males	.793	.709	.831	.964	.558
Females	.846	1.003	.878	.869	.714
Unemployed 14+					
Total unemployed	.818	1.402	.751	.829	.773
Males, white	.739	1.409	.672	.615	.793
Males, black and other	.972	1.180	.828	.922	1.098
Females, white	.816	.921	.933	.765	.710
Females, black and other	.851	1.081	.928	.838	.890
Whites	.757	1.376	.738	.657	.768
Blacks and others	.916	1.159	.874	.874	1.039
Males	.775	1.394	.700	.719	.800
Females	.880	1.012	.917	.893	.760
Employed 16+					
Whites employed in agriculture	.949	1.093	1.090	.727	.968
Whites employed in nonagriculture	.766	1.059	1.045	.693	.664
Blacks and others employed in agriculture	1.082	1.208	1.086	1.081	1.084
Blacks and others employed in nonagriculture	.767	1.022	.959	.699	.923
Whites employed	.659	.976	.697	.695	.547
Blacks and others employed	.800	1.027	.941	.733	.946
Males employed	.800	.765	.814	.938	.594
Females employed	.848	.998	.877	.870	.719

TABLE 2. BETWEEN PSU VARIANCE RATIOS--CONTROLLED SELECTION TO INDEPENDENT SELECTION
(States in the West)

	Arizona	California	Colorado	Hawaii	Idaho	Montana	Nevada	New Mexico	Oregon	Utah	Wash	Wyoming	Alaska
NSR Population 14+	.250	.208	.224	.995	.134	.066	.617	.023	.220	.298	.255	.065	1.109
Enrolled in School 14-34													
Whites	.364	.247	.313	.999	.281	.171	.664	.206	.297	.288	.512	.284	1.081
Blacks and Others	1.079	.617	.700	.998	.562	.985	.742	.886	.686	1.093	.587	.766	1.112
Males	.457	.274	.322	1.000	.268	.175	.669	.168	.308	.302	.508	.351	1.153
Females	.474	.232	.311	.996	.308	.148	.655	.109	.298	.274	.513	.210	1.058
Labor Force Items													
In civilian labor force 14+													
Males, white	.238	.258	.179	.992	.114	.098	.615	.062	.218	.311	.286	.052	1.074
Males, black & other	.945	.451	.588	.993	.566	.979	.656	.855	.578	1.209	.567	.815	1.112
Females, white	.260	.245	.234	.995	.162	.149	.667	.081	.251	.364	.250	.083	1.015
Females, black & other	.993	.445	.644	.997	.493	.971	.691	.849	.567	1.196	.442	.791	1.210
White	.242	.251	.193	.993	.127	.108	.632	.063	.228	.328	.271	.056	1.074
Blacks and others	.963	.428	.587	.995	.515	.977	.667	.852	.572	1.209	.498	.802	1.163
Males	.238	.259	.180	.993	.113	.086	.615	.035	.217	.320	.284	.048	.997
Females	.273	.242	.239	.996	.162	.137	.665	.049	.251	.374	.246	.081	.964
Unemployed 14+													
Total unemployed	.238	.293	.218	.999	.380	.172	.707	.079	.238	.280	.376	.096	.968
Males, white	.243	.325	.211	.996	.475	.216	.744	.141	.246	.257	.392	.107	1.054
Males, black & others	1.010	.732	.660	.990	.915	1.249	.819	.914	.684	1.042	.807	1.075	1.087
Females, white	.218	.280	.288	1.007	.314	.213	.693	.181	.282	.352	.397	.120	1.059
Females, black & others	.876	.637	.954	1.005	.692	.972	.820	.821	.686	.942	.626	.854	.988
Whites	.216	.295	.211	1.002	.381	.182	.720	.113	.242	.278	.377	.093	1.065
Blacks and others	.973	.683	.736	.998	.830	1.182	.819	.873	.662	.996	.752	1.019	1.045
Males	.335	.326	.211	.992	.475	.232	.731	.120	.240	.261	.398	.119	1.009
Females	.195	.275	.307	1.006	.312	.181	.683	.140	.279	.351	.388	.122	1.094
Employed 16+													
Whites employed in agriculture	.492	.821	.484	.998	.384	.547	.996	.478	.366	.632	.341	.465	1.043
Whites employed in non-agriculture	.245	.243	.196	.993	.158	.166	.642	.060	.239	.352	.279	.069	1.067
Blacks and others employed in agriculture	.589	.622	.821	1.014	.633	.967	1.000	.799	.799	1.124	.785	1.040	1.083
Blacks and other employed in nonagriculture	.997	.435	.591	.992	.531	.936	.676	.855	.572	1.236	.427	.687	1.198
Whites employed	.245	.250	.195	.993	.126	.111	.629	.064	.228	.333	.267	.057	1.070
Blacks and other employed	.959	.411	.575	.994	.491	.924	.668	.852	.579	1.231	.461	.727	1.183
Males employed	.238	.257	.183	.993	.113	.092	.612	.039	.218	.327	.282	.050	1.016
Females employed	.283	.243	.240	.996	.159	.140	.666	.049	.251	.378	.237	.084	.950