VARIANCE ESTIMATION FOR STATE ESTIMATES FROM THE EXPANDED CURRENT POPULATION SURVEY

Lawrence Cahoon
U.S. Bureau of the Census

I. INTRODUCTION

The Comprehensive Employment and Training Act of 1973 (CETA) provides for the allocation of Federal funds to prime sponsors within the individual States on the basis of the "relative number of unemployed persons within the State as compared to such numbers in all States." [11]

At the request of the Bureau of Labor Statistics (BLS), the Census Bureau designed an expansion to the Current Population Survey (CPS) to produce State estimates that meet the reliability requirements of BLS. This required the selection of additional sample in approximately half the States.

This paper presents a general overview of the proposed variance estimation procedure for those States where an additional sample was chosen. These are two main areas of interest. These are the use of a collapsed stratum variance estimator and the use of a weighted average of sample data and census data variance estimators.

The usual procedure when a collapsed stratum variance estimator is used is to form each collapsed stratum from two of the original strata. Consideration is given to forming collapsed strata containing two or more of the original States. This procedure is evaluated from a mean-square error viewpoint.

The use of a census data variance estimate is considered due to the few numbers of sample areas in each State. While such a variance estimate is correct only at the time of the census, the estimate is being proposed in order to reduce the mean square error of the final variance estimate.

In order to facilitate the discussion of these two areas, a brief description of the CPS design and the design of the supplemental sample are given in Section II. A more complete description of the CPS design is given in [12]; a more complete description of the supplemental sample is given in [3]. Also discussed in Section II are some initial considerations in the variance estimation and aspects of the variance estimation which are preliminary to the two main issues discussed in this paper. The discussion of the collapsed stratum variance estimation, the methods for forming the collapsed strata, and the evaluation of these three methods are contained in Section III. The discussion of the two proposed census data variance estimators and the reasons for their use is given in Section IV.

II. INITIAL CONSIDERATIONS

A. The CPS Sample Design

Under the current CPS design the United States is divided into 1,924 primary sampling units (PSU's). These PSU's are grouped into 376 strata. One hundred and fifty-six of these strata contain only one PSU; the PSU's in these strata are included in the sample with certainty and are designated as self-representing (SR). The remaining PSU's are grouped in 220 strata with each stratum containing more than one PSU; the PSU's in these strata are designated as nonself-representing (NSR). The creation of these strata was done with the intention of obtaining the best national estimates and thus strata frequently cross State lines. In each stratum containing NSR PSU's a single PSU was selected with probability proportionate to size. Additionally, the 220 strata are grouped into 110 pairs. From each pair one stratum is selected at random (i.e., with equal probability). One PSU was chosen from the selected stratum with probability proportionate to size. Selection of the PSU's was independent for the two procedures.

B. The Sample Design in the States Where A Supplemental Sample was Chosen

In each of the States where additional sample was necessary in order to obtain the required degree of reliability on the State estimates, a supplemental sample, referred to as the CETA sample, was designed which attempts to maximize the use of the national CPS sample. Those PSU's which are self-representing in the CPS national design are retained as self-representing in the CETA State design.

The CETA design as it relates to the NSR PSU's is more complex. A requirement of the CETA design was that all NSR PSU's within a State be represented by a sample PSU within the State. The CPS strata cross State lines; therefore, at the first stage of the CETA design the in-State portion of each CPS national stratum was defined to be a CETA State stratum. These State strata were than divided into two groups. The first group contains those State strata which do not contain a sample PSU. The PSU's within these strata were regrouped into a new set of State strata. A single sample NSR PSU was selected within each stratum with probability proportionate to size. The second group of State strata were those which do contain sample PSU's. The State strata contained in this second group are retained and no additional sample PSU's are chosen within these State strata.

As a result of this procedure, a different selection of national CPS sample PSU's would have generated a different set of CETA State strata and a different set of CETA sample PSU's. Thus the strata definitions for the CETA design are random events. Nevertheless, the procedure was such that overall probabilities of selection were determinable and the resulting sample unbiased.

The estimation procedure for the CETA sample is similar to that used for the national CPS sample. A simple unbiased estimate is prepared by multiplying the value for each characteristic for each sample unit by the probability of selection of the sample unit. A noninterview adjustment by State is

made next to account for nonresponse. A first-stage ratio estimate is then produced by State, based on 1970 census totals, to adjust for differences in population characteristics in the sample PSU's and in the entire State. A national second-stage ratio adjustment is then made to the sum of the State first-stage ratio estimates based on the age, sex, race distribution of the United States population.

The discussion which follows focuses on the variance estimator of the unbiased estimate of population totals. The results presented may be extended to estimates produced at the suceeding stages of estimation without major modifications.

C. The Effects of the Random Strata Definitions on the Variance

The creation of the redefined strata was dependent upon the CPS strata in the States which are represented by national CPS sample PSU's in the in-State portion of each CPS stratum. This resulted in strata definitions for CETA in the supplemental States being random events.

We can express the variance over all possible samples for an estimate, Y, of the population total for a given characteristic as

$$\mathrm{Var}(Y) = E_1\,(\mathrm{Var}_2(Y)) + \mathrm{Var}_1(E_2(Y)).$$

The condition variance, $\mathrm{Var}_2(Y)$, and the conditional mean, $E_2(Y)$, are evaluated over all possible samples given a fixed strata definition. E_1 and Var_1 are evaluated over the range of possible strata definitions. We focus first on the term $\mathrm{Var}_1(E_2(Y))$. As was indicated in Section II-B, the expected value of the unbiased estimate over all possible samples given any fixed set of strata is a constant: i.e., $E_2(Y)$ is a constant. Thus $\mathrm{Var}_1(E_2(Y))$ is zero. $\mathrm{Var}_2(Y)$ is the variance of the sample estimate if the strata definitions were not random events; over all possible samples $\mathrm{Var}_2(Y)$ is an unbiased estimate of $E_1(\mathrm{Var}_2(Y))$. We propose to estimate $E_1(\mathrm{Var}_2(Y))$ in the usual fashion by $\mathrm{Var}_2(Y)$.

D. Variance Estimation in the Self-Representing PSU's

The only component of variance in the SR PSU's is the within-PSU variance. This variance will be estimated in the same manner as is done for the SR strata in the CPS design with a few minor modifications.

E. Variance Estimation in the Strata Containing NSR PSU's - An Introduction

The primary problem encountered in variance estimation for the estimates from the strata containing NSR PSU's is that there are a relatively few number of such strata in each supplemented State. This makes it difficult to obtain a variance estimate which can be regarded as reliable. The design in these strata which resulted from the CETA supplementation to the CPS meant that, with only a few exceptions, each stratum is represented by sample from a single PSU.

1. Estimation of the Within-PSU Variance in the NSR PSU's.

   The within-PSU component of variance for the NSR PSU will be estimated in exactly the same manner as for SR PSU's.

2. Estimation of the Total Variance for the Strata Containing NSR PSU's.

   The estimation of the total NSR variance is to be the weighted average of three variance estimates. The first estimate will be obtained by means of sample data using a collapsed stratum variance estimate. The second and third estimates will utilize the sample data estimate of within-PSU variance and two different estimates of between-PSU variance obtained using Census data. These three variance estimates are described in greater detail in the following sections.

III. Estimation of Total NSR Variance from Sample Data

A. Introduction and Theory

Hansen, Hurwitz and Madow [4] give the following formula for a collapsed stratum variance estimate:

$$\sum_{g}^{G} \frac{L_g}{L_g - 1} \sum_{h}^{L_g} \left(x'_{gh} - \frac{A_{gh}}{A_g}\, x'_g\right)^2 \qquad (1)$$

where A_{gh} is the population of stratum h in group g, x'_{gh} is the estimate for stratum h in group g, A_g is the population of group g, and x'_g is the estimate for group g.[1]

Normally L_g is taken as two if a collapsed stratum variance is to be used. The research discussed below was undertaken to determine the optimum size for L_g and whether the size of the groups can be varied to obtain "better" estimates of the variance for the sample which resulted from the CETA design. Methods are developed for approximating the variance of the estimator and the bias in the estimator. These approximations are then used to define groups into which the strata are placed. Three States are considered in detail. As a result of the research it was decided, with a few exceptions, to place all strata containing NSR PSU's in a single group for all supplemented States. As we are primarily interested in an accurate measure of the variance of the unemployment estimate, the evaluations utilize unemployment data.

It can be shown that the collapsed stratum variance estimate, formula (1), tends to be an over-estimate of the true variance. Hansen, Hurwitz and Madow [4] show that formula (1) is a biased estimate; specifically the expected value of formula (1) is

$$\sum_{g}^{G}\sum_{h}^{L_g}\sigma^2_{x'_{gh}} + \sum_{g}^{G}\left\{\frac{1}{L_g - 1}\left[V^2_{A_{g(h)}} - 2V_{A_{g(h)}}\sigma^2_{x'_{g(h)}}\right]\sum_{h}^{L_g}\sigma^2_{x'_{gh}}\right\}$$

$$+ \sum_{g}^{G}\frac{1}{L_g - 1}\sum_{g}^{L_g}\left(X_{gh} - \frac{A_{gh}}{A_g}X_g\right)^2 . \qquad (2)$$

where

$$V_{A_{g(h)},\sigma^2_{x'_{g(h)}}} = \frac{\sum_{h}^{L_g} A_{gh}\sigma^2_{x'_{gh}}}{\bar{A}_g\,\sigma^2_{x'_g}} - 1 , \qquad (3)$$

with

$$\sigma^2_{x'_g} = \sum_{h}^{L_g}\sigma^2_{x'_{gh}} , \quad \bar{A}_g = \frac{1}{L_g}\sum_{h}^{L_g} A_{gh} \qquad (4)$$

and where

$$V^2_{A_{g(h)}} = \frac{\sum_{h}^{L_g} A^2_{gh}}{L_g\bar{A}^2_g} - 1. \qquad (5)$$

The second and third terms in expression (2) are the biases in the collapsed stratum variance estimate. Expression (5) can be recognized as the relvariance of the strata sizes. We wish to approximate the bias in the estimated variance depending on the size of the groups and the composition of the groups.

In order to simplify the evaluation of the bias in the variance estimates, assume that the total variance for a stratum is proportional to the size of the stratum. That is, assume

$$\sigma^2_{x'_{gh}} = cA_{gh}.$$

This is a good assumption for national CPS and for the supplemented States in CETA where the between-PSU variance is a small proportion of the total variance. Under this assumption equation (3) becomes

$$V_{A_{g(h)},\sigma^2_{x'_{g(h)}}} = \frac{\sum_{h}^{L_g} A^2_{gh}}{L_g\bar{A}^2_g} - 1 = V^2_{A_{g(h)}}$$

Thus the second term in expression (2) becomes

$$-\sum_{g}^{G}\frac{1}{L_g - 1}V^2_{A_{g(h)}}\sum_{h}^{L_g}\sigma^2_{x'_{g(h)}} .$$

Thus the bias for group g from the second term in expression (2) expressed as a percent of the variance for group g is

$$\frac{100}{L_g - 1}\left[1 - \frac{\sum_{h=1}^{L_g} A^2_{gh}}{L_g\bar{A}^2_{gh}}\right] \text{ percent.} \qquad (6)$$

Since this term is dependent only on the size of the strata we may properly designate it as the bias due to differences in stratum sizes within group g. Unless the strata vary widely in size this term tends to be small.

The third term in expression (2) is the bias due to differences in the characteristics of the strata. This term can be determined directly from census values; however, those calculations are correct only for one point in time. We will assume, with some caution, that those calculations will be indicative of the present magnitude of this term. While we can approximate the magnitude of this term from census data, to obtain an estimate of the relative bias due to this term we must know the total NSR variance of the estimate of level. We can approximate this value.

The total variance for a group of strata for unemployment items can be approximated by $b_s X_g$ where $X_g = \sum_h X_{gh}$ is the census value for the item for group g and b_s is the product of the State NSR design effect and the State NSR sampling interval.[2] This is an acceptable approximation to the NSR variance as the number of unemployed persons is a small percentage of the total population. The computation of the design effects for States is based on the work by C. Dippo using 1960 census data with one change. Based on data obtained since the CETA supplementation we have assumed a 1.1 within-PSU design effect instead of the 1.4 within-PSU effect used by C. Dippo [3].[3]

The approximate percent bias for each group due to differences in the characteristic across strata can then be approximated by:

$$\frac{100\ L_g}{L_g - 1}\sum_{h=1}^{L_g}\left(X_{gh} - \frac{A_{gh}}{A_g}X_g\right)^2 \Big/ b_s X_g \qquad (7)$$

The numerator of expression (7) is the bias within group g and is obtained from the third term of expression (2). The denominator is the approximate variance in the State for group g based on the above design effect.

B. Three Methods of Forming Collapsed Strata

For the convenience of our discussion we will describe three methods for grouping the strata within the State and associate an estimate with each of these methods. The three methods are:

Method I. Place all strata in a single group.

Method II. For this method all strata are placed in groups of size two (i.e., $L_g = 2\ V_g$); if there is an odd number of strata, one group consists of three strata. The pairing of strata for the group is done so as to have strata with similar size and characteristics in the same group. This is done so as to minimize the bias in the variance estimate.

Method III. For this method the strata are placed in groups of varying size, the only constraint being that all strata in a given group be of similar size and have similar characteristics. Methods I and II are special cases of this method of grouping.

C. Comparison of the Bias for the Three Methods

We will consider three States as examples in the computation of the biases. Subsequently, we will compute an approximate mean square error for the estimates discussed here. The characteristic of interest in these evaluations is the 1960 census unemployment level.

1. Arkansas. The population, the 1960 census unemployment rates, and the projected unemployment rates for the strata in Arkansas are given in table 1. The projected unemployment rate was obtained from the step-wise regression program used to determine the strata definitions for the supplemented States in the CETA expansion [3].

Table 1 ARKANSAS
Stratum Unemployment Characteristics

Stratum	In-State 1970 Population	Projected 1970 Unemployment Rate	1960 Unemployment Level
577	131501	0.0183	1555
590	88882	0.0282	2494
757	158725	0.0227	2656
658	76859	0.0216	1402
681	124112	0.0176	2006
914	87124	0.0190	1881
945	149436	0.0222	2617
AR1	110080	0.0154	1924
AR2	126966	0.0209	2582
AR3	112639	0.0226	2847
AR4	122101	0.0244	2571
AR5	125169	0.0304	2859
Total	1413597	--	27394

The tabulations in table 1 were used to form the groups and evaluate the biases for each of the three methods. For Arkansas, differences in stratum population were not considered in forming the group since the bias resulting from these differences is small. To obtain an estimate of the bias for Method I, all strata were placed in a single group. For Method II, a more complicated procedure was used. First, it was felt that the groups should be formed based upon the 1970 projected unemployment rates since these rates were used to form the strata in the State and, second, we wished to have strata with similar characteristics in the same group. Therefore, the groups were formed according to the following procedure. The two strata with the lowest projected unemployment rate formed the first group. The next group contained the two strata with the lowest projected unemployment rates among the strata remaining. This procedure was continued until six groups of two strata each were formed. This procedure results in the minimum bias possible among all possible groupings with two strata per group when the bias is computed based on the 1970 projected unemployment rates.

In forming the groups for Method III we used the same considerations as were used for Method II. Again we wished to have strata with similar characteristics in the same group. We again used the projected unemployment rates to form the groups. Two constraints were imposed in forming the groups. First, at least one group had to contain more than two strata; this prevented Methods II and III from resulting in the same set of groups. The second constraint was that the projected unemployment rate for all strata in group i be less than the projected unemployment rate for each stratum in group j if $j > i$. These constraints allow for several different groupings of the strata in the State. Each of these were considered as a possible grouping for the strata. Once these constraints have been satisfied there are several possible sets of groups for Method III. It was decided to choose the grouping which satisfied the given constraints and which minimized the relative mean square error of the variance estimate when the bias is computed based on the 1960 census unemployment data. This methodolgy actually gives an unfair advantage to Method III because the characteristic of interest is used to determine the stratification. The methodology used to estimate the relative mean square error of the variance estimate is described below. For each of the groupings resulting from the three methods formula (2) can be used to estimate the total NSR variance. For the State of Arkansas the groups resulting from each of these three methods are:

Method I. A single group of all twelve NSR strata.

Method II. Six groups -- (AR1,681), (577,914), (AR2,658), (945,AR3), (657,AR4), (590,AR5).

Method III. Three groups -- (577,681,AR1), (914,658,AR2,945,657,AR3,AR4), (590,AR5).

The estimates of the relative bias in the variance estimates when each of these three sets of groups are used to estimate the variance are given in table 2.

Table 2 ARKANSAS
Bias in the Estimates of Variance

	Method I.	Method II.	Method III.
Degrees of Freedom	11	6	9
Bias Due to Differences in Stratum Population	-0.36%	-2.74%	-1.16%
Bias Due to Differences in Stratum Characteristics	6.70%	6.12%	3.64%
Net Bias	6.34%	3.38%	2.48%

The interesting result from Table 2 is that the use of the collapsed stratum variance estimation procedure does not minimize the bias when each group contains only two strata.

2. South Dakota. The data used to form the groups for the Strata in South Dakota for each of the three methods is given in table 3. The resulting biases for the three methods are given in table 4.

Table 3 SOUTH DAKOTA

Stratum	In-State 1970 Population	Projected 1970 Unemployment Rate	1960 Unemployment Level
464	48336	0.0087	813
SD1	31269	0.0069	440
SD2	34485	0.0091	441
SD3	31748	0.0098	554
SD4	30930	0.0099	505
SD5	36587	0.0104	539
SD6	40338	0.0115	699
SD7	32348	0.0119	377
SD8	37718	0.0126	348
SD9	30965	0.0129	693
SD10	31632	0.0157	621
SD11	30277	0.0292	748
TOTAL	415633	--	6778

Table 4 SOUTH DAKOTA
Bias in the Estimates of Variance

	Method I.	Method II.	Method III.
Degrees of Freedom	11	6	10
Bias Due to Differences in Stratum Population	-0.20%	-1.33%	-0.20%
Bias Due to Differences in Stratum Characteristics	5.21%	6.84%	2.67%
Net Bias	5.01%	5.51%	2.47%

The groups for South Dakota are:

Method I. A single group of all 12 NSR strata.

Method II. Six groups of two strata each -- (SD1, 464), (SD2,SD3), (SD4,SD5),(SD6,SD7), (SD8,SD9), (SD10,SD11).

Method III. Two groups -- (SD1,464,SD2,SD3,SD4, SD5,SD6,SD7,SD8),(SD9,SD10,SD11)

3. Idaho. For Idaho the bias in the variance estimate is extremely large regardless of which of the three methods is used. This illustrates the need to utilize a census data variance estimator as a part of the total variance estimate. The results of the bias calculation are given in table 6. The tabulations used to form the groups and to compute the approximate bias are given in table 5. The groupings for each of the three methods are:

Method I. One group containing all six strata.

Method II. The groups are (ID1,840), (ID2,807), (ID3,ID4).

Method III. ID1,840,ID2), (807,ID3,ID4).

Table 5 IDAHO

Stratum	In-State 1970 Population	Projected 1970 Unemployment Rate	1960 Unemployment Level
807	48205	0.0237	856
840	55151	0.0099	832
ID1	49129	0.0094	837
ID2	50089	0.0143	603
ID3	47983	0.0242	790
ID4	53943	0.0494	2329
TOTAL	304499	--	6247

Table 6 IDAHO
Bias in the Estimates of Variance

	Method I.	Method II.	Method III.
Degrees of Freedom	5	3	4
Bias Due to Differences in Stratum Population	-0.06%	-0.27%	-0.29%
Bias Due to Differences in Stratum Characteristics	53.31%	46.37%	43.93%
Net Bias	53.25%	46.10%	43.64%

The large biases observed in table 6 in the variance estimators for Idaho are due to the large differences between stratum ID4 and the other strata in the State. Stratum ID4 has twice the unemployment rate of any other NSR stratum in the State; this difference remained unchanged between 1960 and 1970. This example illustrates that we should not blindly use the collapsed stratum variance estimation technique; rather, we should do a careful evaluation of the procedure on a State-by-State basis.

D. The Variance of the Variance Estimates

We wish to evaluate the methods based on the mean square error of the variance estimate. First we must approximate the the variance of the variance estimate.

Hansen, Hurwitz and Madow [4] give the following formula for the relvariance of the estimated variance when proportionate stratified sampling is used and when the strata are of equal size and the within stratum variances are all the same.

$$Z^2 = \frac{1}{n}\left(\bar{\beta} - \frac{\bar{n}-3}{\bar{n}-1}\right), \qquad \bar{\beta} = \frac{\sum_{h=1}^{L} \mu_{4h}}{L\bar{s}^4}$$

where μ_{4h} is the fourth central moment of stratum h, L is the number of strata, $\bar{s}^2$ is the within stratum variance, $\bar{n}$ is the number PSU's selected per stratum and n is the total number of PSU's selected. The assumption made in this formula are restrictive and are not met exactly in the CETA sample. However, the assumptions are not so restrictive that formula (8) cannot be used as an approximation to the relvariance of the estimate of variance. In using formula (8) we will make the additional assumption, that μ_{4h} is the same for all strata. Thus the expression for $\bar{\beta}$ reduces to $\bar{\beta} = \mu_4/\bar{s}^4$. Thus $\bar{\beta}$ is the kurtosis of the within stratum distribution for the characteristic. Since we are using a sample where the variates take on the values 1 or 0 (i.e., 1 if unemployed and 0 if

employed) we have a binomal distribution and

$$\bar{\beta} = \frac{1}{PQ} - 3 \qquad (9)$$

where P is the proportion of people with the characteristic. For Method I the relvariance of the variance estimate is

$$Z_2^2 = \frac{1}{n}\left(\bar{\beta} - \frac{n-3}{n-1}\right) \qquad (10)$$

For Method II, where $\bar{n} = 2$, the relvariance is

$$Z_1^2 = \frac{1}{n}(\bar{\beta} + 1) \qquad (11)$$

The relvariance of the estimate from Method III cannot be defined from formula (8). For this reason, we approximate its relvariance for Method III, Z_3^2, by linear interpolation based on the degrees of freedom for each method.

These approximations can than be used to approximate the relative mean square error, Rel-MSE, of the estimates.

We are primarily interested in obtaining accurate estimates of the variance of yearly averages. We do know that for unemployment items the variance of a yearly estimated average is approximately 20 percent of the variance of a monthly estimate. Based on this, we assume the same relationship for the variance of the variance estimate. Large variations from the factor of five rarely influences the choice of methods.

E. Comparisons of the Mean Square Error of the Three Methods

Utilizing this assumption and the theory previously developed, tables 7 and 8 present the Rel-MSE for each of the three methods for Arkansas and South Dakota respectively.

Since the major concern is with estimating State unemployment and variance of that estimate values of P, the percent of the population unemployed, between 0.03 and 0.05 are of primary interest. The Rel-MSE for these values of P are given in Tables 7 and 8. It can be shown from equations (12) and (13) that which method minimizes the Rel-MSE is not dependent upon the value of P.

Table 7 ARKANSAS
Relative Mean Squared Error of the Estimates of Variance

P = 0.03	Rel-MSE Monthly Estimate	Rel-MSE Yearly Average	Degrees of Freedom
Method I	2.5492	0.5131	11
Method II	2.6978	0.5405	6
Method III	2.6064	0.5218	9
P = 0.05			
Method I	1.4400	0.2912	11
Method II	1.5886	0.3186	6
Method III	1.4972	0.2999	9

Table 8 SOUTH DAKOTA
Relative Mean Squared Error of the Estimates of Variance

P = 0.03	Rel-MSE Monthly Estimate	Rel-MSE Yearly Average	Degrees of Freedom
Method I	2.5477	0.5116	11
Method II	2.6997	0.5424	6
Method III	2.5761	0.5157	10

Table 8 (cont'd)

P = 0.05	Rel-MSE Monthly Estimate	Rel-MSE Yearly Average	Degrees of Freedom
Method I	1.4385	0.2897	11
Method II	1.5705	0.3205	6
Method III	1.4669	0.2939	10

The calculations of the previous two sections are relative to only unemployment and do not take into account other characteristics. Except for some minor considerations used in Method III the calulations do not allow for changes in the characteristics of the strata over time. It is felt that changes in the characteristics of the strata over time should in general affect Method II the most and Method I the least. On the basis of minimum relative mean square error. Based on this criterion, we choose Method I for Arkansas and North Dakota. The magnitude of the bias in the estimates for Idaho indicated that special consideration should be given to that State. In most of the remaining supplemented States a similar analysis indicates that Method I is to be preferred.

The actual method chosen does not make a large difference in the relative-mean-square error of the variance estimator. The frequently used procedure is to form groups of size two, i.e., use Method II. This data indicates that while there is little difference among the methods, Method II is the worst of the three.

IV. Estimation of Total NSR Variance from Census Data

The estimation of the total NSR variance from census data, in fact, utilizes the census data to estimate only the between-PSU component of variance. The within-PSU component of variance will be estimated from sample data as outlined in Section II-D.

The first census data estimate of between-PSU variance takes the usual form of the variance over all possible samples. This variance estimator is:

$$C^2 \left(\sum_h \sum_i \frac{P_h}{P_{hi}} X_{hi}^2 - \sum_h X_h^2 \right)$$

where

P_h is the 1970 census population in stratum h, P_{hi} is the 1970 census population in PSU i in stratum h, X_h is the 1960 census total for the characteristic for stratum h, X_{hi} is the 1960 census total for the characteristic for PSU i in stratum h, $C = \hat{u}/(\sum_h X_h)$, and $\hat{u}$ is the current survey estimate for the characteristic.

The term C^2 is included in expression (8) to adjust the variance estimate for the differences in the level of the estimate between 1960 and the time of the survey.

The second census data between-PSU variance estimate is not the typical variance estimate. Instead, it is a direct measure of the squared error due to the given selection of sample PSU's within the NSR strata. The use of this variance estimator was suggested by Gary Shapiro of the Bureau of the Census. The variance estimate is

$$C^2 \left[\sum_h \left(\frac{P_h}{P_{hi}} X_{hi} - X_h \right) \right]^2$$

Here i is the PSU in stratum h which is in sample. This formula provides the best measure of the actual squared error resulting from the selection of the NSR PUS's. It accounts for the difference between the characteristics of the PSU's in sample and the characteristics of the State taken as a whole. Thus this form of variance estimator is specific to the actual set of PSU's selected whereas formula (8) is not. Formula (8) provides a variance estimate over all possible samples.

Current plans are that the final variance estimate be a weighted average of the sample data and the two census data variance estimates. The within-PSU variance is to be estimated entirely from census data. The census data variance estimate will be used to reduce the mean square error of the between-PSU variance estimate. The weights to be used have not been determined.

REFERENCES

[1] Cochran, W.G., "Sampling Techniques," John Wiley and Sons, New York (1963).

[2] Dippo, C., "CPS-CETA Documentation: Strata Listing for States Supplemental for CETA," Internal Census Bureau Document, February 25, 1976.

[3] Dippo,C., "Expansion of CPS to Provide Reliable State Estimates of Unemployment," Proceedings of the American Statistical Association, August, 1975.

[4] Hansen, M.H., Hurwitz, W.N., and Madow, W.G., "Sampling Survey Methods and Theory," Vol. I, Vol. II, John Wiley and Sons, New York (1953).

[5] Keyfitz, Nathan, "Estimates of Sampling Variance When Two Units are Selected from Each Stratum," Journal of the American Statistical Association, 52 (1957) pp. 503-510.

[6] McCarthy, P.J., "Pseudo-Replication: Half Samples," Review of the International Statistical Institute 37; Number 3 (1969) pp. 239-264.

[7] Shapiro, Gary M., "Keyfitz Method of Estimating Variance and Its Application to the Current Population Survey," Internal Census Bureau Document, September 1, 1966.

[8] Thompson, M.M., Shapiro, G., "The Current Population Survey: An Overview," Annals of Economic and Social Measurement, Vol. 2, No. 2 (1973).

[9] Tepping, B.J., "Variance Estimation in Complex Surveys," Proceedings of the American Statistical Association, August 1968.

[10] Woodruff, R.S., "Simple Method of Approximating the Variance of a Complicated Estimate," Journal of the American Statistical Association, June 1971, pp. 411-414.

[11] _____, "Comprehensive Employment and Training Act of1973," Public Law 93-203, 93rd Congress, S-1559, December 28, 1973.

[12] _____, "U.S. Bureau of the Census, "The Current Population Survey--A Report on Methodology," Technical Paper Number 7, 1963, (This paper is currently being revised).

[1] A_{gh} and A_g need not be population totals but may be the value of any known characteristic correlated with Ex'_{gh} and Ex'_g.

[2] This is a slight deviation from our earlier assumption that stratum variance is proportional to the size of the characteristic. The assumptions are approximately equivalent.

[3] Based on more recent data the within-PSU des design effect is being revised. Current indications are that the design effect is between 1.3 and 1.5.