Estimating Sampling Variance from the Current Population Survey:

A Synthetic Design Approach to Correcting Standard Errors

Dean Jolliffe[*]

Economic Research Service, U.S. Department of Agriculture

Paper presented at the 2002 meeting of the
American Agricultural Economics Association, Long Beach, CA

Abstract: Essentially all empirical questions that are addressed with sample data require estimates of sampling variance. The econometrics and statistics literatures show that these estimates depend critically on the design of the sample. The sample for the U.S. Current Population Survey (CPS), which serves as the basis for official poverty, unemployment, and earnings estimates, results from a stratified and clustered design. Unfortunately, analysts are frequently unable to estimate sampling variance for many CPS statistics because the variables marking the strata and clusters are censored from the public-use data files. To compensate for this, the Bureau of Census provides a method to approximate the sampling variance for several, specific point estimates, but no general method exists for estimates that differ from these cases. Similarly there are no corrections at all for regression estimates. This paper proposes a general approximation method that creates synthetic design variables for the estimation of sampling variance. The results from this method compare well with officially reported standard errors. This methodology allows the analyst to estimate sampling variance for a significantly wider class of estimates than previously possible, and therefore increases the usefulness of research resulting from the CPS data files.

Classification: C81, I32, O18
Key Words: Current Population Survey, Sampling Variance, Sample Design, Regional Analysis, Rural, Poverty

I. **Introduction**

Essentially all empirical questions that are analyzed using survey data require estimates of sampling variance. Without estimates of sampling variance, there are no credible answers to whether some point estimate is different from zero, has changed over time, whether it varies for persons with different characteristics, or whether there are rural and urban differences in the point estimate. This is a rudimentary point that is well understood throughout the economics literature, but the issue of constructing estimates of the sampling variance that approximate the true population variance of some point estimate may be difficult for many surveys.

The standard econometric and statistics textbooks used by economists provide estimates of sampling variances for frequently used statistics such as means and regression parameter estimates, but they typically assume that the sample data are drawn following a simple random design. See for example, DeGroot (1986), Greene (2000), Johnston and DiNardo (1997). Most nationally representative data sets, though, are not simple random draws, but are typically stratified and/or multi-stage sample designs. As one example, the sample used for the U.S. Current Population Survey (CPS), which serves as the basis for official poverty, unemployment, and earnings estimates, is drawn from a census frame using a stratified, multi-stage design.

Estimating sampling variance for complex sample designs is an issue that is also well understood and documented in the statistics literature. Kish (1965), Cochran (1977), and Levy and Lemeshow (1999) all derive estimates for the sampling variance of estimated means, ratios, and other descriptive statistics for a wide variety of complex sample designs. Similarly, Holt *et al*. (1980), Nathan and Holt (1980), Scott and Holt (1982) provide estimates of the sampling variance for regression parameters for complex sample designs and Wu *et al*. (1988) provide an estimate of the variance of the F statistic under a two-stage design.

There are several examples illustrating that estimates of the sampling variances will be severely downward biased if it is erroneously assumed that data come from a simple random draw when in fact a complex, multi-stage sample design was used. Howes and Lanjouw (1998, Table 2) present evidence that estimated standard errors for the Foster-Greer-Thorbecke (1984) poverty indices are dramatically effected by sample design. In their examples, the design-corrected standard errors for the poverty indices are between 26 and 56 percent larger when the design effects are accounted for.

1

Deaton (1997, Table 1.5) uses data from the stratified, multi-stage Pakistan Integrated Household Survey and shows that the estimated standard errors for the national average household expenditure increases by 72 percent when correcting for the sample design effects. Scott and Holt (1982) use data from the U.K. General Household Survey and the Family Expenditure Survey to provide examples showing that the estimated variance of OLS parameter estimates may also increase dramatically under a multi-stage sample design.[1]

In all of these examples the authors not only had access to the unit-level data records, but the data files also contain information indicating the survey strata and clusters. Indeed the literature on sampling assumes that the analyst can identify the strata and clusters of the sample in order to estimate the sampling variance. In the case of the CPS data, all information on sample design has been censored from the data files, so the analyst can not directly estimate sampling variance.[2]

To compensate for the missing sample design data from the CPS public data files, U.S. Bureau of Census (2000, Appendix C) provides detailed notes on how to approximate design-corrected standard errors for a limited set of labor, income and poverty estimates. The recommended way of estimating standard errors is essentially based on parameterizing the relationship between a direct estimate of the design-corrected variance and the relevant statistic. This method, called the generalization method, results in two coefficients (a, b) which can be used to approximate the design-corrected variance based on the point estimate and sample size.

An important shortcoming of the generalization method is that the (a,b) coefficients are only provided for a limited set of point estimates. This constrains analysis along several dimensions. First it constrains the types of variables that can be analyzed. For example, following the Census methodology it is possible to estimate the design-corrected standard error for total income, but not for expenditure or savings or receipt of government transfers. Along these same lines, it is possible to estimate the sampling variance for the number of unemployed but not for the number of discouraged or part-time workers.

---

[1] Similarly Deaton (1997, Table 2.1) uses data from rural Pakistan to show that the t-values from an ordinary least squares regression of commodity unit values on expenditure and household size also decrease significantly (in one example they are reduced by more than one half).

[2] The purpose of censoring the data is to maintain confidentiality of the data, as it is assumed that researchers would be able to pinpoint individual respondents if the strata and PSU information were included.

The second way in which the methodology constrains the researcher is that it limits the types of geographic and demographic comparisons that can be made for each of the parameter estimates. For example, Appendix G of U.S. Bureau of Census (2000) provides coefficients to estimate the variance for the estimated number of persons in the labor force, in agricultural employment, or unemployed. Coefficients are also provided for subsamples of each of these estimates by race, ethnicity and gender. Using this method it is possible to construct a design-corrected estimate of the variance for the number of White women in agricultural employment, but, for example, no coefficients exist to estimate the variance for the number of Black men in agricultural employment.

A third way in which the methodology limits research questions is that it only provides corrections for estimates that are sums, percentages, differences, or ratios. There is no correction provided for regression parameters, so all regression analysis from the CPS public data files most likely are reporting t-statistics that are significantly biased upwards. As another example of this limitation, there are no corrections provided for any measure of poverty other than the head count index. The methodology does not allow for the estimation of standard errors for any distribution-sensitive measure of poverty.

In this paper I propose an alternate method of estimating sampling variance which addresses the shortcomings of the recommended methodology. The method is based on creating synthetic variables that possess characteristics similar to what is known about the CPS sample design. Estimates of the confidence intervals derived from this synthetic approach are compared to estimates reported in Bureau of Census documents (Dalaker and Proctor, 2000) and those derived from the recommended approximation method. The results show that the estimated confidence intervals compare well with the officially reported estimates.

The primary advantage of the synthetic approach is that it can be applied to a wide variety of point estimates and for any subsample of the population. In addition, the synthetic approach can be used for several alternate measures of poverty or for any mean value including conditional means from regression analysis. The plan of this paper is as follows. Section II provides an overview of the implications of stratification and multi-stage sampling for estimating sampling variance. Section III describes in more detail how the generalization method for estimating standard errors is implemented and discusses the implications of this approach with a particular focus on nonmetropolitan estimates. This section continues with a discussion of the CPS sample design and a

description of the derivation of the synthetic design variables. Section IV evaluates the performance of the synthetic approach by comparing the estimates with officially reported estimates of 90% confidence intervals for a wide variety of point estimates. Section V provides a brief conclusion.

II. Overview of Sample Design Concepts

There are primarily two features of a sample design, stratification and clustering, which distinguish a simple random design from a complex design. In a simple random design the ultimate sampling units, such as firms, households, or individuals, are randomly drawn in one take, typically by mapping a list of random numbers to a complete list of the reference population. In a stratified design, the list of the reference population, or the sample frame, is first explicitly divided (typically either geographically or on some demographic characteristic) into smaller sections. Then from each of these smaller sections, called strata, the selection process continues by randomly drawing a fixed proportion of the sample from each stratum.

In a multi-stage design, the ultimate sampling units are not selected in one draw from the sample, but rather are the result of stages of drawing. Again the sample frame is split into several sections, this time they are called clusters or primary sampling units. After the frame is divided into clusters, the first-stage of the selection process proceeds by randomly selecting clusters.[3] In a two-stage design, once the clusters are selected, the ultimate sampling units are randomly drawn from each of the selected clusters. Lengthier multi-stage designs are also frequently used where after selection of the clusters, secondary sampling units are drawn, then possibly tertiary sampling units, and so on, before the ultimate sampling units are drawn.

Stratification and Sampling Variance

Both stratification and clustering have implications for the estimated variance which are most easily understood when their effect is contrasted to the sampling variance resulting from a pure random sample. Generally it is the case that stratification reduces the sampling variance. Kish (1965) illustrates this by

---

[3] One type of random selection process is for clusters to be selected with probability proportional to the estimated population of the cluster, though there are other designs such as a randomly selecting clusters with equal probabilities. The decisions made on the type of design have implications for the efficiency of the design and for the resulting sampling weights.

first noting that the estimated variance of a stratified sample is equal to the stratum-weighted average of the variances of each stratum, or:

$$Var(\overline{x}) = 1/n \sum_h w_h \sigma_h^2 \tag{1}$$

where $\overline{x}$ is the sample mean, strata are denoted by the subscript h, σ_h^2 is the within-strata variance, w_h is the stratum weight, and n is the sample size. This result follows since the draws from each stratum are independent samples.

In order to examine the relationship between the variance from a stratified sample and the variance resulting from a simple random sample (srs), it is useful to express the deviation of each element from the sample mean as the deviation of the element from its stratum mean plus the deviation of the stratum mean from the sample mean:

$$(x_{h,i} - \overline{x}) = (x_{h,i} - \overline{x}_h) + (\overline{x}_h - \overline{x}) \tag{2}$$

After squaring and summing both sides of this expression, one obtains:

$$\sum_h \sum_i (x_{h,i} - \overline{x})^2 = \sum_h \sum_i (x_{h,i} - \overline{x}_h)^2 + \sum_h n_h (\overline{x}_h - \overline{x})^2 + 2 \sum_i (x_{h,i} - \overline{x}_h) \sum_h (\overline{x}_h - \overline{x}) \tag{3}$$

where n_h is the sample size of stratum h. If all terms are divided by the sample size, n, then the left-hand side of this expression is an expression for the sample variance under the assumption that the sample was drawn from a simple random sample. The first term on the right-hand side is the variance when the sample is stratified, and the last term on the right-hand side of this expression sums to zero. This can also be expressed as:

$$Var(\overline{x}_{stratified}) \approx Var(\overline{x}_{srs}) - \sum_h w_h (\overline{x}_h - \overline{x})^2 \tag{4}$$

The last term on the right-hand side is positive and expresses the reduction in the sampling variance from stratifying the sample. One implication of this is that the greater is the heterogeneity across stratum, the greater is the efficiency gain to stratification. Unusual strata increase sampling efficiency, but the gain is proportional to the strata weights, so a highly unique stratum will not increase efficiency by much if it only represents a small portion

of the population. Another implication of the expression is that the more homogeneous observations are within each stratum, then the greater is the relative efficiency gain from stratification.

The expression also shows, though, that the relative efficiency gain will be small if the population variance is large. This may be true even when the sum of the squared deviations of the strata means from the population mean is large. Kish (1965, Section 4.6) asserts that it is frequently difficult to find strata such that the squared deviations of strata means from the population means are large relative to the population variance; and therefore Kish states that the gains over simple random sampling are often not that large.

Cluster or Multi-stage Selection

The benefit of a simple random sample is that each sample observation is independently drawn. This is in contrast to the multi-stage design where once a cluster is selected the draws within that cluster are typically not independent because observations within a cluster are frequently more similar to each other than observations drawn across clusters. This clustering of observations frequently induces positive correlation between the cluster elements and this often results in a large increase in sampling variance.[4]

To illustrate the loss of estimation efficiency from clustering relative to a simple random design, I consider a two-stage design and follow an example from Deaton (1997). Consider some variable x, where

$$x_{i,c} \quad = \quad \mu \, + \, \alpha_c \, + \, \varepsilon_{i,c} \tag{5}$$

where i subscripts the ultimate sampling unit, say household or individual, and c subscripts the cluster. The mean of $x_{i,c}$ is μ, α_c is the cluster effect, and $\varepsilon_{i,c}$ is a random variable with a mean of zero and variance of σ_ϵ^2. The distribution of $\varepsilon_{i,c}$ is independent and identical for all i and c. Similarly, α_c is a random variable with mean of zero, variance of σ_α^2 and is independently and identically distributed across all c, and is also independent of ϵ. The independence assumptions come from the random draws of clusters and then the random draws of households within clusters. These assumptions imply:

$$E(\overline{x}) = \mu \quad \text{and} \quad V(\overline{x}_{2s}) = \frac{\sigma_\alpha^2}{n} + \frac{\sigma_\varepsilon^2}{nm} \tag{6}$$

---

[4] The benefit of multi-stage designs is that they can dramatically reduce the costs of the survey field work by reducing the travel costs associated with interviewing each observation.

where n is the number of clusters, m is the number of households in each cluster, and the subscript 2s denotes that the sample is a two-stage design. By collecting terms over a common denominator, then adding and subtracting σ_α^2, and finally by multiplying and dividing by $\sigma_\alpha^2 + \sigma_\varepsilon^2$, $V(\overline{x}_{2s})$ can be rewritten as:

$$V(\overline{x}_{2s}) = \frac{\sigma_\alpha^2 + \sigma_\varepsilon^2 + (m-1)\sigma_\alpha^2}{nm} = \frac{\sigma_\alpha^2 + \sigma_\varepsilon^2}{nm} + \frac{(\sigma_\alpha^2 + \sigma_\varepsilon^2)(m-1)\sigma_\alpha^2}{nm(\sigma_\alpha^2 + \sigma_\varepsilon^2)} \tag{7}$$

By recognizing that the variance resulting from a simple random sample is given by $[\sigma_\alpha^2 + \sigma_\varepsilon^2]/nm$ and by defining the coefficient of intra-cluster correlation, ρ, as $\sigma_\alpha^2/[\sigma_\alpha^2 + \sigma_\varepsilon^2]$, then the relationship between the sampling variance from a two-stage design and a simple random sample can be expressed as:

$$V(\overline{x}_{2s}) = V(\overline{x}_{srs})[1+\rho(m\text{-}1)] \tag{8}$$

where the subscript srs denotes simple random sample.[5] From (8) it is clear that the size of the cluster and the intra-cluster correlation coefficient are the determinants of the correction for the design effect. From (8) it can also be seen that if one were to incorrectly assume that some sample design were a pure random sample and estimated the variance of $\overline{x}$, then if the true design were a two-stage design, the true sampling variance would be underestimated by a factor of $[1+\rho(m-1)]$. This factor is also sometimes denoted as *deff,* and is called the design effect.

The implications of (8) are that the estimation efficiency loss from the two-stage design becomes greater as the number of households per cluster, m, increases (holding total sample size fixed). This implication can be easily seen by noting that if m is equal to one, then the design is a simple random draw. Similarly the efficiency loss from the cluster design becomes greater as the intra-cluster correlation coefficient, ρ, increases. One way to understand this is to consider some variable that is exactly the same for everyone in the cluster, such as a whether some public facility exists in that cluster. In this case, it is not correct to assume that the sample contains m independent draws of this variable from a particular cluster. Once it is determined that for one observation in a cluster the facility exists, then this is known to be true for all observations in the cluster.[6] The estimation

---

[5] This derivation assumes clusters are of equal size. For a more general derivation, see Kish (1965).
[6] Moulton (1990) provides a very useful illustration of this case.

implications for this are that the effective sample size for the facility variable is not n, but rather n/m or the number of clusters in the sample.

Deaton (1997, Table 1.5) finds in his examples from the Pakistan Integrated Household Survey (PIHS) that the corrections for stratification are very small, less than 0.5 percent for the mean values he considers. In contrast, from the same table Deaton shows that the effect of clustering on the estimated sampling variance can be very large. As one example the PIHS data show that the estimated cluster-corrected standard error for per capita expenditure is 42 percent greater than the incorrect standard error based on assuming that the data come from a simple random sample. Kish and Frankel (1974) note:

> In stratification negative correlation reduces the variance; but that gain is less for subclass means, and even less for their differences and for complex statistics. Clustering induces larger and positive correlations between element values. The resulting increase in variance is measured by the ratio *deff*, and is often severe. [p.1].

III. Sampling Variance and the Current Population Survey

Because sample design information is censored from the public CPS data files, direct estimation of design-corrected sampling variance is not possible. Users of the CPS public data files can approximate the sample variance using parameter estimates that are provided in CPS user's guides, such as U.S. Bureau of Census (2000). In order to derive these parameter estimates, the Bureau of Census first directly estimates the sampling variance for a variety of point estimates. The direct estimates of the variance of various point estimates are found through a replication method that is similar in principle to a bootstrap methodology. A series of random subsamples are selected from the original sample in a manner that replicates the sample design. From each of these subsamples the relevant point estimate is produced, and the series of estimates result in an empirical estimate of the distribution of the point estimate. This empirical estimate of the distribution of the point estimate can be used to directly estimate the sampling variance of the point estimate.

Once direct estimates of the variance for several different point estimates are derived, then generalized variance functions are estimated of the form:

$$\text{Var}(X) = aX_i^2 + bX_i + \varepsilon_i \tag{9}$$

where X is some sample estimate and ε is the error term. The functional form for equation (9) is based on the assumption that the distribution of the random variable is approximated by the binomial distribution. U.S. Bureau of Census and Bureau of Labor Statistics (2000) show that the variance of the sample estimate is approximately equal to the product of the design effect and the variance of the sample estimate assuming a simple random sample and a binomial distribution. An implication of this distributional assumption is that the coefficient b is equal to the product of the design effect and the population raising factor (ratio of population to sample size), or b=*deff**N/n.

The Bureau of Census estimates equation (9) through an iterative weighted least squares procedure which reduces the influence of sample estimates with large relative variances. The resulting (a, b) coefficients are reported for a variety of statistics and can be used to approximate design-corrected standard errors with the public CPS data files. See U.S. Bureau of Census and Bureau of Labor Statistics (2000) for more details.

Appendix G of U.S. Bureau of Census (2000) and Appendix C of Dalaker and Proctor (2000) reports these coefficients for several types of estimates and for subsamples of several different characteristics. The shortcoming of this approach is that it is not possible to provide coefficients for all relevant estimates, and the analyst may therefore have to reshape their research question in order to fit the categories provided or proceed without correcting the standard errors for design effects.[7] For example, using the (a,b) coefficients provided in Dalaker and Proctor, the analyst can examine the incidence of poverty for persons between, say, the ages of 15 and 24 but not for working-age adults or teenagers. Similarly following the Census methodology it is possible to estimate the standard error for the total number of White males or Black males, but it is not possible to estimate the standard error for the number of White or Black males with income less than some specified amount.

If the relevant point estimate happens to match the categories provided in the CPS user manuals or other supporting documents, then estimating the design-corrected standard error is straightforward. For example, U.S. Bureau of Census (2000) shows that the correct standard error for a estimated percentage is given by:

$$s_{x,p} = \sqrt{\frac{b}{x}p(100-p)} \tag{10}$$

where p is the estimated percentage and x is the total number of weighted observations in the base of the percentage. In the case of the estimated standard error for a percentage, it is assumed that the denominator is a population statistic, not an estimated sample statistic. If, the denominator is a sample statistic, then one is really estimating a ratio of two sample statistics, and the standard error for an estimated ratio, x/y, is given by:

$$s_{x/y} = \frac{x}{y} \sqrt{\left(\frac{s_x}{x}\right)^2 + \left(\frac{s_y}{y}\right)^2 - 2r\frac{s_x s_y}{xy}} \tag{11}$$

where $s_x = \sqrt{ax^2 + bx}$ and r is equal to the correlation between x and y. U.S. Bureau of Census (2000) state that the analyst should assume r is equal to zero.[8]

Metropolitan and Nonmetropolitan Estimates

In addition to the shortcomings already discussed of using equations such as (10) and (11) to estimate standard errors, there is another issue to consider for those who are examining differences between metropolitan and nonmetropolitan statistics. In particular, for the labor characteristics such as the number of persons employed or unemployed, no coefficients are provided to examine these estimates for metropolitan areas separately from nonmetropolitan areas.

For estimates related to the number of people by demographic characteristics or poverty status, several table notes in U.S. Bureau of Census (2000, Appendix G) and Dalaker and Proctor (2000, Appendix C) state that the (a,b) coefficients for metropolitan areas are the same as for the entire sample, and those for nonmetropolitan areas are all multiplied by 1.5. It is difficult to understand, though, why it is that these coefficients are all increased by a factor of 1.5. Recalling that the theoretical motivation for the general variance functions implies that the coefficient b is equal to the design effect times the raising factor, or *deff**N/n. The implication of the footnote then is that:

---

[7]As the examples from Deaton illustrate, proceeding without correcting for the design effects results in a dramatic underestimate of the true standard error.

[8] There is one stated exception to this and that is when the denominator is the count of families, the numerator is the number of persons in each family with some characteristic, and at least one person has this characteristic in each family. Census provides an example of this as the number of children per family for those families with children. In this case, the analyst is to assume that r is equal to 0.7.

$$b_{\text{nonmetro}} = 1.5 * b_{\text{metro}} \qquad \Rightarrow \qquad \frac{deff_{\text{nonmetro}}}{deff_{\text{metro}}} = 1.5 * \frac{N_{\text{metro}}}{n_{\text{metro}}} * \frac{n_{\text{nonmetro}}}{N_{\text{nonmetro}}} \qquad (12)$$

One striking implication of equation (12) is that for all estimates over the sample of metropolitan and nonmetropolitan samples, the ratio of the design effects is constant. Given the large variation that one observes in intracluster correlation for various characteristics, it would seem that the adjustment by 1.5 for the nonmetropolitan coefficients results in estimates of the sampling variance that are less precise than for other estimates.

It is not possible to test this implication with the public use data as one can not obtain the estimated design effects independent of the (a,b) coefficients, but in Table 1 below I explore the credibility of this assumption by treating other geographic variables as if they were the cluster variables. In the first column, the 50 states (plus the District of Columbia) are treated as the cluster variable, and design effects are estimated by nonmetropolitan and metropolitan areas. Similarly, in the next two columns, counties are treated as the cluster variables. For the smallest counties there is also a concern about revealing the identities of individual observations so many of these observations are aggregated into one large county within each state. The first "County" column includes these aggregated counties, and the second column excludes them.

[INSERT TABLE 1 APPROXIMATELY HERE]

The estimated design effects presented in Table 1 all assume that the average metropolitan and nonmetropolitan cluster size, m, is equal to four. Following Gleason (1997), intracluster correlation, as defined in equation (7), is estimated as:

$$\rho = \frac{\dfrac{\sum_i w_i (\bar{x}_i - \bar{x})^2 (n-k)}{\sum_i \sum_j w_{ij} (x_{ij} - \bar{x}_i)^2} - (k-1)}{\dfrac{\sum_i w_i (\bar{x}_i - \bar{x})^2 (n-k)}{\sum_i \sum_j w_{ij} (x_{ij} - \bar{x}_i)^2} + \sum_i w_i - \sum_i w_i^2 / \sum_i w_i - (k-1)} \qquad (13)$$

where i subscripts each of the k clusters, j subscripts the observations within each cluster, n is the sample size, $\bar{x}$ is the sample mean, $\bar{x}_i$ is the mean of cluster i, w_{ij} is the weight for each observation, and w_i is the sum of cluster i weights.

The assumption that the nonmetropolitan (a,b) coefficients are equal to 1.5 times the metropolitan coefficients means that the ratio of these design effects should be constant over all variables.[9] For the exercise displayed in Table 1, I examine this ratio where I used the state and county variable information to proxy for the cluster variable. From these results it is clear that there is a lot of variation in the ratios which suggests that the '1.5-rule' is a fairly rough adjustment for analyzing nonmetropolitan characteristics.

In addition to assuming that the appropriate adjustment for nonmetropolitan estimates is the same over all estimates, there appears to also be the assumption that this adjustment doesn't change over time. The adjustment factor of 1.5 for nonmetropolitan estimates was first suggested for CPS data from 1981 in U.S. Bureau of Census (1983, Table B-7) and has not changed over the years. This is in marked contrast to the coefficients for other estimates listed in the appendices to the CPS user manuals (e.g. U.S. Bureau of Census, Appendix G, 2000) or the Census P-60 Series on Poverty (e.g. Dalaker and Proctor, Appendix C, 2000) that are updated annually. Personal communication with Bureau of Census sheds a bit more light on the development of this adjustment factor:[10]

> The factor of 1.5 has been used for non-metropolitan areas as a simple approximation. While the best factor likely varies from characteristic to characteristic, we use 1.5 for all characteristics rather than publishing a different factor for each estimate. Years ago, someone looked at the data for metro/non-metro areas and decided that 1.5 would be a good, and somewhat conservative, estimate for most characteristics.

Because the adjustment for nonmetropolitan areas is not specific to different characteristics and since it is not frequently updated, it is reasonable to assume that Census has decided that providing a full set of updated adjustment factors for nonmetropolitan characteristics is a lower priority. The analyst examining nonmetropolitan characteristics needs to be aware of this when using 1.5-adjustment factor.

---

[9] That is the ratio should be constant over all variables where the metropolitan and nonmetropolitan samples considered are unchanged.

[10] I'm grateful to Alfred Meier from the Bureau of Census for this explanation, as well as for his help with many inquiries related to estimating sampling variance from the CPS data.

Synthetic Design Approach

The idea of the approach proposed in this paper is to mimic certain aspects of the actual design of the CPS sample by creating synthetic variables for the strata and clusters which induce similar design effects. From equation (8) it is apparent that the two characteristics of the cluster design that have the greatest effect on sampling variance are the cluster size and the magnitude of intracluster correlation. The ultimate sampling units (USUs) for the CPS are drawn from the geographically sorted primary sampling units (PSUs or clusters) following a fixed-interval, systematic selection procedure. In essence each housing unit within a PSU is numbered, a random number is used to select the first housing unit, and then each of the next housing units selected is determined by the size of the fixed interval. U.S. Bureau of Census and Bureau of Labor Statistics (2000, p. 3-7) state, "…most USUs consist of a geographically compact cluster of approximately four addresses, corresponding to four housing units at the time of the census."

The first step of the synthetic design approach is to choose the variable of primary interest and sort the data file on this variable. Table 2 examines poverty, and so I sort the data by income. If a researcher is looking at expenditure or savings, then the first step is to sort on these variables. The next step of the design approach is to assign each consecutive four housing units to a separate cluster.[11] The purpose of the sorting is to induce a high level of intracluster correlation, and the choice of four matches the average cluster size of the CPS.

Having created the synthetic clusters, the next step is to select the strata. Section II of this paper notes the statements from Kish and Frankel and the example from Deaton to support the assertion that clustering has a relatively greater effect than stratification in terms of estimating sampling variance. Following this assertion, the synthetic design approach proposed in this paper is more closely linked to the actual clustering aspect of the CPS design than the stratification.

From equation (4) it can be seen that the main efficiency gain from stratification comes from choosing strata that have differences in means across strata and that are more homogenous within strata.[12] The CPS strata are geographically contiguous and are selected to insure that they are "as homogeneous as possible with respect to

---

[11] The CPS variable name for the household sequence number is fh_seq and this is used to identify unique households.

labor force and other social and economic characteristics that are highly correlated with unemployment" (U.S. Bureau of Census and Bureau of Labor Statistics, 2000, p.3-2). To capture the geographic aspect of the stratification, I select as the synthetic strata the four regions of the United States (Northeast, Midwest, South and West). These regions are selected because there are significant mean differences across these regions with respect to labor force and other social and economic characteristics that are correlated with unemployment.

With the selection of the synthetic strata and clusters one can then directly obtain design-corrected estimates of sampling variance. As examples, Kish (1965) provides design-corrected estimates of sampling variance for sample means and other basic descriptive statistics, and Scott and Holt (1982) provide design-corrected estimates of the sampling variance for ordinary least squares estimates. As a brief review of the literature on these corrections, recall that the estimated sampling variance of the sample mean from a weighted, simple random sample (srs) is given by:

$$\hat{v}(\overline{x}_{w,\,srs}) = n(n-1)^{-1} \sum_{i=1}^{n} \left(w_i \Big/ \sum_{k=1}^{n} w_k \right)^2 (x_i - \overline{x}_w)^2 \tag{14}$$

where w_i is the weight for observation i and $\overline{x}_w$ is the weighted mean. Then the estimated sampling variance of the sample mean from a weighted, stratified, clustered sample is given by:

$$\hat{v}(\overline{x}_{w,\,2s}) = \sum_{h=1}^{L} n_h (n_h - 1)^{-1} \sum_{i=1}^{n_h} \left(\sum_{j=1}^{m_{h,i}} w_{h,i,j}\, x_{h,i,j} \;-\; \sum_{i=1}^{n_h} \sum_{j=1}^{m_{h,i}} w_{h,i,j}\, x_{h,i,j} \right)^2 \tag{15}$$

where the h subscripts each of the L strata, i subscripts the cluster or primary sampling unit (PSU) in each stratum, j subscripts the ultimate sampling unit (USU), so x_{hij} denotes element j in PSU i and stratum h. The number of PSUs in stratum h is denoted by n_h, and the number of USUs in PSU (h, i) is denoted by m_{hi}.

IV. Results

Table 2 presents information on poverty in the U.S. for 1999 using the 2000 CPS files, and replicates a large part of Table A in Dalaker and Proctor (2000). Each row presents a poverty estimate for a separate geographic or demographic category, that are grouped into classifications such as family status, race, age, region, and family

---

[12] In contrast to the importance of the number of USUs per cluster, equation (4) indicates that the number of strata is not directly related to the effect of stratification on the estimated sampling variance.

characteristics. The first column lists the estimated percentage or ratio of poor persons in each category.[13] The next four columns provide varying estimates of the 90 percent confidence interval each derived from the Bureau of Census report on poverty (Dalaker and Proctor, 2000). The second to last column lists the estimated 90 percent confidence interval from using the synthetic design approach and equation (15). The last column lists the estimated 90 percent confidence interval from erroneously assuming a pure random sample with weights and then using equation (14) to estimate the sampling variance.

The first estimate of the 90 percent confidence interval is labeled as "Reported in Table A" and is directly copied from the report by Dalaker and Proctor. Bureau of Census explains that these confidence intervals are derived by first estimating standard errors, rounding these to one decimal point, then multiplying by 1.645, and again rounding to one decimal point. As one example, the estimated standard error for the proportion of poor "Black Families" is 0.86. This is rounded to 0.9 and then multiplied by 1.645 to give 1.48, which is then rounded to 1.5.[14]

The second estimate of the 90 percent confidence intervals, labeled "Implied by Levels", also comes from the same report. In addition to reporting the proportion of persons or families that are poor, Table A of this report also reports the 90 percent confidence interval in terms of the number of poor persons. This estimate of the confidence interval in terms of the number of poor persons implies a confidence interval in terms of the percentage poor.

The next column lists equation (10) as the header and reports the 90 percent confidence interval as estimated by methodology recommended by the Bureau of Census for estimating sampling variance for percentages. The fourth estimate of the 90 percent confidence interval, labeled "a,b Ratio", is derived from equation (11) following the recommended methodology for estimating sampling variance for ratios.

The effective distinction between a percentage and a ratio is that when referring to a percentage, Census means that the denominator is a population statistic (and not a sample estimate) whereas a ratio implies that the numerator and denominator are both sample statistics (with estimates of sampling variance). The rule used to

---

[13] Recall from equations (10) and (11) that the semantic difference between ratio and percentage matters.
[14] The description of the reported estimates and the example are from personal, email communication with Alfred Meier of the Bureau of Census. Dr. Meier also notes that in future reports, only the estimate of the 90 percent confidence interval will be rounded, so that there will not be two stages of rounding which creates greater error.

distinguish the two is "if the denominator is a person level characteristic to which we [Bureau of Census] control (i.e. age, sex, race, Hispanic/Non-Hispanic origin, and state) then it will not have a standard error" and it is considered a percentage, otherwise all other estimates are ratios.[15]

This rule seems to imply, for example, that all of the poverty estimates by race and age categories are treated as percentages and those related to types of families are ratios. Nonetheless, the documentation on this issue is sparse and analysts may well be uncertain whether to use equation (10) or (11). For the sake of comparison I include both in Table 2, and have bolded the estimates based on whether Census considers the estimate to be a percentage or ratio (as confirmed through personal communication). When using the (a,b) methodology, it is important to note that this method results from regressing several direct estimates of the sampling variance on similar point estimates, as described in Section III and equation (9). This method minimizes a weighted square of errors, but the resulting estimates of sampling variance are approximations (containing error) to the direct estimates.

The decision of which demographic characteristics to list in the first column of Table 2 is guided by a desire to compare confidence intervals from the synthetic method with official estimates either reported in or derived from Dalaker and Proctor (2000). For many of these characteristics, though, there is not a direct match in the description of the (a,b) estimation approach (Dalaker and Proctor, 2000, Appendix C). The column labeled "Match a,b Categories" lists whether Bureau of Census provides (a,b) coefficients for the estimates reported in Table A. In those cases where no match exists, I attempt to find a category that seems similar, as I assume many analysts might also take this approach. For example, there are no coefficients to estimate the sampling variance for the poverty level of non-Hispanic Whites, so I use the (a,b) coefficients for Whites. While this attempt to match categories is sometimes a reasonable approach, it is important to note that it is likely to introduce further error into the estimates of the sampling variance.

The issue of what values to use for the (a,b) coefficients is even more difficult when the estimate of interest covers two separate categories for which (a,b) coefficients are provided. For example, suppose that the analyst is interested in the sampling variance of the poverty estimate for persons over the age of 45. Since this is considered

---

[15] This explanation is from personal communication with Alfred Meier of the Bureau of Census.

a percentage by Census, the analyst uses equation (10) and only needs a parameter estimate for b. Appendix C of Dalaker and Proctor shows that the value of b is 3,927 for individuals between the ages of 45 and 64, and it is the same for those 65 years of age and older. Given that the estimates are the same, an analyst might assume that the appropriate value of b for those over 45 is also 3,927; but this would be wrong. The fact that the b value is the same for each category implies that there is some similarity in the variance for each age category, but it doesn't imply anything about the variance across the two age categories. If there were large differences in poverty levels across the two age groups, then the total variation of the combined age groups will be greater than the within-group variation and the appropriate b-value is higher.

The purpose of showing four different estimates for the confidence intervals that are either directly from or implied by the Bureau of Census (Dalaker and Proctor) report on poverty, is to show that the official estimates appear to vary substantially. The difference across the estimates results from several sources including the rules for rounding followed by Census, whether the estimate is treated as a percentage or ratio, and whether there is a direct match for the assigned (a,b) coefficients. The four estimates are also meant to illustrate that it is not straightforward to decide which estimate is the best approximation. Where there is a direct match in (a,b) categories, the best approximation is given by equations (10) or (11) depending on whether the analyst considers the estimate to be a percentage or a ratio. When there is no direct match in the (a,b) categories, choosing the best estimate is difficult.

As a result of this assertion that the baseline for comparison is not always clear, I consider the range of design-corrected estimates derived from the Dalaker and Proctor report as a guide for comparison. The estimates reported in the column labeled "Synthetic Design" result from the synthetic clusters and strata, which are created as described in Section III, and following equation (15). As a lower bound for comparison, in the last column I include the 90 percent confidence interval from using equation (14), which correctly accounts for the weights but erroneously ignores the strata and clusters.

Perhaps the most important comparison to consider, is that all of the design-corrected estimates (including the synthetic corrections) are typically much higher than the estimates resulting from assuming a simple random sample. In the case of the national estimate of the proportion of persons poor, the design effect is greater than

17

four. This implies that if the analyst only accounts for the weights of the CPS design, but ignores the cluster and strata design effects, the estimated standard errors will underestimate the correct standard error by more than 100 percent.

Across the 37 poverty estimates listed in Table 2, the synthetic design approach to estimating sampling variance performs very well. In particular, if one compares the estimates from the synthetic design approach with those listed in Table A of the Dalaker and Proctor report, there are several cases where the design approach outperforms the official estimates. As one example, consider that the officially reported estimate of the confidence interval for the national poverty level is reported as 0.3.[16] Once the analyst learns, though, that Census considers this estimate a percentage, then it is possible to estimate this more precisely as 0.33 by using the (a,b) methodology and equation (10).[17] Table 2 shows that the estimate from the synthetic cluster approach is 0.33 also. While all of these estimates round to 0.3, there will be many questions that require more than the one-digit of significance provided by the official estimate of 0.3.[18]

As another example of where the synthetic design approach performs as well as the official estimate, recall the example of the poverty level of Black families. In this case, the officially reported confidence interval is 1.5 and the estimate from the synthetic design approach is 1.32. This would appear to be a case where the synthetic design approach has not performed well. Even in this case, though, the synthetic approach does better than is immediately apparent. The official estimate of 1.5 results from taking the estimated standard error of 0.8606 and rounding this to 0.9 and using this estimate to find the confidence intervals. If one more properly used the estimated standard error of 0.8606 to find the confidence interval, and then rounded, the result would be 1.4. In this case, due to the two stages of rounding, the official estimate is too large by about the same amount that the synthetic estimate appears to be too small.

---

[16] See the first row of Table 2 and the column labeled "Reported Table A".

[17] Note that in this example, there exist values for the (a,b) coefficients for the national poverty estimates.

[18] A retort to the assertion that 0.33 is more useful than 0.3 would be to suggest that I'm ascribing too much precision to the estimate of 0.33. Verifying whether this is true would require a careful analysis of the private CPS files to construct estimates of the dispersion of the sample estimates of the variance. This is not possible in this report, but I assume that the sampling variance can be reasonably reported to more than one digit of significance. Further, it is also important to recall that the official estimate of 0.3 results from estimating the standard error, rounding to one digit, multiplying by 1.645, and again rounding to one

To summarize the results of the 37 poverty estimates provided in Table 2, there are 24 instances where the confidence interval found through the synthetic design approach falls into the range of estimates derived from the Dalaker and Proctor report. Correspondingly, there are 13 cases where the estimate from the synthetic approach is either less than the lowest of the four estimates derived from the Dalaker and Proctor report or greater than the highest estimate. Of these 13 cases, though, there are only four cases where the synthetic design approach falls outside of the range of official estimates by more than ten percent. It is also noteworthy that one of these four estimates where it appears the synthetic approach has performed poorly is for nonmetropolitan areas. Given that these estimates are based on rule of adjusting the (a,b) coefficients by a factor of 1.5 and that this factor has not been updated in the last twenty years, it is not at all clear that the reported estimates for nonmetropolitan characteristics are more accurate than the synthetic-design estimates.

[INSERT TABLE 2 APPROXIMATELY HERE]

In order to compare the performance of the synthetic-design correction for regression estimates, Table 3 examines three regressions of poverty status on race and region. For the case where the binomial variable of poverty status is regressed on dummy variables for race, the regression coefficients will be the same as the percent poor for each race and the regression standard errors will be the same as the standard errors of the sample means.[19] The relationship between these regressions and the sample means can be used to show that the synthetic-design correction performs as well for regression analysis as it does for analysis of means.

Table 3 shows that the synthetic-design correction for the regression estimates matches the corrections for the sample means, which supports the assertion that this method is a useful tool for regression analysis also. This claim is fully consistent with the results from Scott and Holt (1982) who

---

digit. This is a process that certainly introduces a greater level of error in the estimates relative to the expected sampling dispersion of the variance estimates.

[19] This assumes that the constant is suppressed, that the race dummies account for all observations and that there is no overlap in the race dummy variables.

show that the sample-design correction for regression estimates is conceptually very similar to the correction for sample means.

In addition to comparing the synthetic-design correction for the regression estimates with the design-corrected estimates for the sample means, Table 3 also examines the uncorrected OLS standard errors. The purpose of this comparison is based on the assumption that most regression analysis done with CPS data has fully ignored the issue of correcting the sampling variance for design effects. While these regressions exaggerate the importance of this correction, it is sobering to note that all of the regression standard errors in Table 3 increase by more than 80 percent when correcting for stratification and clustering.

[INSERT TABLE 3 APPROXIMATELY HERE]

V. Conclusion

In this paper, I propose a general methodology for estimating sampling variance for estimates from the U.S. Current Population Survey that creates synthetic design variables and then uses the standard formulas from the statistics literature to estimate standard errors. I show that this methodology performs well by examining poverty levels by numerous categories. In particular, for those instances where it is not possible to utilize the (a,b) methodology recommended by Census, the results from Table 2 show that the synthetic methodology performs quite well in approximating design-corrected confidence intervals and is certainly a far superior approach when the alternate is to ignore the design effects.

There are four advantages to this methodology over the methodology suggested by the Bureau of Census. First, it is possible to examine a wider class of variables. For example, following the Census methodology the analyst can estimate the sampling variance for estimates related to income, but no corrections are provided for expenditures, or savings, or receipt of government transfers. Second, for a given variable, it is possible to examine more tabulations (or 'breakdowns') by relevant demographic characteristics with the synthetic methodology. Census provides approximations if the analyst wishes to examine, for example, the number of poor persons 15

years of age and older, but no approximations are provided if the analyst wishes to examine a category not specified by Census, such as number of adults that are poor.

The third advantage of this methodology is that it allows the analyst to estimate sampling variance for a wider class of point estimates. Census provides approximations for means, ratios, and percentages of several different variables, but if the analyst is interested in different point estimates no corrections are provided. One example of this is in terms of poverty indices. Dalaker and Proctor (2000) provide adjustments for the headcount index, but not for any other poverty index, such as the poverty gap, or squared-poverty gap index. Similarly, and with broader implications, it is not possible to correct regression estimates using the Census methodology, but the synthetic approach allows this. The fourth advantage is simply that the methodology is relatively easy to implement. The analyst needs to create appropriate synthetic variables and then there are several statistical softwares that will then estimate design-corrected standard errors.[20]

All of these advantages are important, but this paper does not argue against using the Census-recommended methodology at all. In those cases where the researcher is examining a variable (and relevant demographic subcategories) for which Census has provided factor adjustments for estimating the sampling variance, then the researcher is best advised to follow the Census methodology. This recommendation is moderated somewhat when considering a nonmetropolitan variable. In this case the adjustments have not been updated over the last twenty years and they are likely to be significantly less useful than other adjustments provided by Census.

For any of the cases described above where the variable of interest, or demographic breakdown, or parameter estimate varies from those for which Census provides adjustments, this synthetic approach provides the research with an alternative methodology of estimating sampling variance. This methodology allows the analyst to estimate sampling variance for a significantly wider class of estimates than previously possible, and therefore increases the usefulness of research resulting from the CPS data files.

---

[20] For example, the following softwares all provide estimates of design-corrected standard errors: Bascula from Statistics Netherlands, CENVAR from U.S. Bureau of the Census, CLUSTERS from University of Essex, Epi Info from Centers for Disease Control, Generalized Estimation System (GES) from Statistics Canada, IVEware (beta version) from University of Michigan, PCCARP from Iowa State University, SAS/STAT from SAS Institute, Stata from Stata Corporation, SUDAAN from Research Triangle Institute, VPLX from U.S. Bureau of the Census, and WesVar from Westat, Inc.

REFERENCES

Cochran, William G. *Sampling Techniques.* New York: Wiley Series in Probability and Statistics, 1977.

Dalaker, Joseph and Proctor, Bernadette. "Poverty in the United States: 1999." *U.S. Census Bureau, Current Population Reports,* Series P60-210, Washington, DC, 2000.

Deaton, Angus. *The Analysis of Household Survey*. Baltimore: John Hopkins, 1997.

DeGroot, Morris. *Probability and Statistics*. Second Edition, Reading: Addison-Wesley Publishers, 1986.

Foster, James; Greer, Joel and Thorbecke, Erik. "A Class of Decomposable Poverty Measures." *Econometrica*, May 1984, 52(3): 761-765.

Gleason, John. "Computing Intraclass Correlations and Large ANOVAs." *Stata Technical Bulletin,* January 1997, STB-35: sg65.

Greene, William. *Econometric Analysis*. Fourth Edition, Upper Saddle River: Prentice Hall, 2000.

Holt, D.; Smith, T. M. F. and Winter, P. D. "Regression Analysis of Data from Complex Surveys." *Journal of the Royal Statistical Society. Series A (General)*, 1980, 143(4): 474-487.

Howes, Stephen and Lanjouw, Jean Olson. "Does Sample Design Matter for Poverty Comparisons." *Review of Income and Wealth*, March 1998, 44(1): 99-109.

Johnston, Jack and DiNardo, John. *Econometric Methods*. Fourth Edition, New York: McGraw Hill, 1997.

Kish, Leslie, *Survey Sampling*, New York: John Wiley & Sons, 1965.

Kish, Leslie and Frankel, Martin. "Inference from Complex Samples. " *Journal of the Royal Statistical Society, Series B (Methodological),* 1974, 36 (1): 1-37.

Levy, Paul S. and Lemeshow, Stanley. *Sampling of Populations: Methods and Applications.* Third Edition. New York: Wiley Series in Probability and Statistics, 1999.

Moulton, Brent R. "An Illustration of a Pitfall in Estimating the Effects of Aggregate Variables on Micro Units." *Review of Economics and Statistics*, 1990, 72(2): 334-338.

Nathan, G. and Holt, D. "The Effect of Survey Design on Regression Analysis." *Journal of the Royal Statistical Society. Series B (Methodological)*, 1980, 42(3): 377-386.

Scott, Andrew and Holt, Tim. "The Effect of two-stage Sampling on Ordinary Least Squares Methods." *Journal of American Statistical Association*, December 1982, 77 (380): 848-854.

U.S. Bureau of Census. "Current Population Survey: Annual Demographic File, 2000." Inter-university Consortium for Political and Social Research, Document No. 6692, 2000.

U.S. Bureau of Census. "Characteristics of the Population Below the Poverty Level: 1981." Washington, D.C., Current Population Reports, Series P-60, No. 138, 1983.

U.S. Bureau of Census and Bureau of Labor Statistics. "Design and Methodology." Current Population Survey Technical Paper No. 63, March 2000.

Wu, C.F. Holt, D. and Holmes, D.J. "The Effect of Two-stage Sampling on the F Statistic." *Journal of the American Statistical Association*, March 1988, 83(401): 150-59.

Table 1: Ratio of Nonmetro to Metro Design Effects
Treating State, County, and censored County as Cluster Variables

Characteristics	Deff$_{nonmetro}$/Deff$_{metro}$			Characteristics	Deff$_{nonmetro}$/Deff$_{metro}$		
	State	County	County[1]		State	County	County[1]
Number of Persons by gender, race, ethnicity:				*Number of Persons by Income Group:*			
Males	1.00	1.00	1.00	Poor	1.03	0.97	1.05
Females	1.00	1.00	1.00	Less than 10,000	1.02	0.98	0.99
White	1.28	1.12	1.55	10-20,000	1.01	0.99	1.05
Nonhispanic White	1.17	0.99	1.11	20-30,000	1.01	0.98	0.95
Black	1.31	1.13	0.81	60-70,000	1.01	0.99	0.97
Asian	1.34	1.25	0.71	Greater than 70,000	1.02	0.93	0.95
Hispanic	1.09	0.93	0.65				
				# of Families by Type:			
by Education Level:				Married Couple	1.01	0.97	1.04
Less than High School	1.02	0.98	0.97	Female-headed	1.02	1.01	1.08
High School	1.00	0.98	0.99	Female-headed, White	1.01	1.00	0.97
Some College	1.02	1.01	1.02	Female-headed, Black	1.07	1.02	1.03
Bachelors and More	1.01	0.95	0.92	Female-headed, Asian	1.04	1.04	0.96

[1]In the CPS files, small counties are aggregated into one category (geco=0). The first column marked 'County' includes these aggregated counties, and the second column excludes them.
Notes: Estimated design effects for nonmetropolitan and metropolitan areas are approximated by $[1+\rho(m-1)]$, where the cluster size, m, is four and ρ is the intracluster correlation coefficient. State and county variables are used as proxies for the clusters.

Table 2: Comparison of 90% Confidence Intervals for 1999 CPS Poverty Estimates

Characteristics	Ratio or Percent Poor	Direct and Implied Estimates from the P-60 Report					Estimated 90% Confidence Intervals	
		Reported Table A	Implied by Levels	a,b Percentage Equation (10)	a,b Ratio Equation (11)	Match a,b Categories	Synthetic **Design** Equation (15)	Random Sample Equation (14)
Family Status								
Persons	11.8	0.3	0.33	**0.33**	*	yes	0.33	0.16
Persons in Families	10.2	0.3	0.34	0.34	*	no	0.36	0.17
In Unrelated Subfamilies	39.1	4.9	4.27	6.84	**9.15**	no	5.65	3.37
Unrelated Individuals	19.1	0.7	0.62	0.49	**0.66**	yes	0.61	0.51
Unrelated Males	16.3	0.8	0.77	0.66	**0.81**	no	0.78	0.71
Unrelated Females	21.7	1.0	0.87	0.70	**0.94**	no	0.82	0.73
Race								
White	9.8	0.3	0.34	**0.33**	*	yes	0.31	0.16
Non-Hispanic White	7.7	0.3	0.32	**0.32**	0.33	no	0.29	0.16
Black	23.6	1.2	1.20	**1.20**	*	yes	1.24	0.66
Asian	10.7	1.6	1.58	**1.56**	*	yes	1.54	0.83
Hispanic	22.8	1.2	1.23	**1.23**	*	yes	1.05	0.50
Age								
Under 18	16.9	0.7	0.65	**0.65**	0.66	yes/no	0.64	0.37
18-64 years	10.0	0.3	0.39	**0.39**	*	no	0.30	0.20
18-24 years	17.3	0.8	0.78	**0.76**	0.80	no	0.79	0.65
25-34 years	10.5	0.5	0.53	**0.51**	0.54	no	0.49	0.41
35-44 years	8.3	0.5	0.43	**0.43**	0.44	no	0.41	0.35
45-54 years	6.7	0.5	0.43	**0.43**	0.44	no	0.40	0.34
55-59 years	9.2	0.8	0.85	**0.83**	0.88	no	0.81	0.71
60-64 years	9.8	1.0	0.99	**0.95**	1.01	no	0.82	0.76
65 years +	9.7	0.5	0.53	**0.53**	0.53	yes	0.53	0.43

Table continued on next page

Table 2: Comparison of 90% Confidence Intervals for 1999 CPS Poverty Estimates (Continued)

Characteristics	Ratio or Percent Poor	Direct and Implied Estimates from the P-60 Report				Match a,b Categories	Synthetic Design Equation (15)	Random Sample Equation (14)
		Estimated 90% Confidence Intervals						
		Reported Table A	Implied by Levels	a,b Percentage Equation (10)	a,b Ratio Equation (11)			
Region								
Northeast	10.9	0.7	0.70	**0.73**	0.77	no	0.67	0.34
Midwest	9.8	0.7	0.66	**0.63**	0.66	no	0.62	0.32
South	13.1	0.7	0.63	**0.58**	0.61	no	0.62	0.30
West	12.6	0.8	0.77	**0.70**	0.75	no	0.71	0.34
Metropolitan	11.2	0.3	0.36	**0.36**	*	yes	0.36	0.18
Nonmetropolitan	14.2	1.2	1.06	**0.99**	1.07	yes[a]	0.82	0.41
Families								
Total	9.3	0.3	0.33	0.28	**0.34**	yes	0.32	0.29
White	7.3	0.3	0.31	0.27	**0.31**	yes	0.29	0.27
White, Non-Hispanic	5.5	0.3	0.28	0.26	**0.28**	no	0.28	0.27
Black	21.9	1.5	1.35	1.14	**1.42**	yes	1.32	1.28
Asian	10.3	1.6	1.64	1.56	**1.72**	yes	1.66	1.66
Hispanic	20.2	1.5	1.38	1.19	**1.45**	yes	1.02	0.98
Type of Family								
Married Couple	4.8	0.3	0.25	0.23	**0.26**	no	0.25	0.23
Female-headed	27.8	1.5	1.28	1.02	**1.42**	no	1.14	1.07
Female-headed, White	22.5	1.5	1.37	1.17	**1.50**	no	1.19	1.16
Female-headed, Black	39.3	3.0	2.67	2.03	**3.09**	no	2.34	2.29
Female-headed, Asian	23.1	7.4	6.94	5.95	**7.51**	no	6.19	6.19

[a] Nonmetropolitan matches with the adjustment of increasing nonmetropolitan (a,b) coefficients by a factor of 1.5.

Notes: Confidence intervals are listed in percentage points, and the asterisk denotes that the number is undefined (square root of a negative number). The first four columns of confidence intervals are derived from the Dalaker and Proctor (2000) P-60 report on poverty. The bold estimate marks whether Census considers the estimate a percentage or ratio. The next column lists whether there is a direct match in characteristics between the poverty estimates and those characteristics assigned a,b coefficients, described in Section III are listed next, followed by the confidence intervals from assuming that the data are from a weighted, simple random sample.

Table 3: Synthetic Correction for OLS Standard Errors Compared to Uncorrected, OLS Standard Errors and an Analysis of Means, Poverty Regressions and Poverty Means

| | Regression Analysis: Poverty Status on Race and Region | | | Analysis of Sample Means: Poverty Status by Race and Region | | | | | |
Explanatory Variables	Coefficient	OLS Stnd Error	Synthetic Stnd Error	Percent Poor	SRS Stnd Error	Synthetic Stnd Error	Table A Stnd Error	Eq. (2) Stnd Error	Match a,b Categories
Regression 1:									
White	9.8	0.097	0.188	9.8	0.097	0.188	0.182	0.202	yes
Black	23.6	0.401	0.756	23.6	0.401	0.756	0.729	0.728	yes
Asian, Pacific Islander	10.7	0.502	0.932	10.7	0.502	0.932	0.973	0.951	yes
American Indian, Eskimo	28.7	1.368	2.541	28.7	1.368	2.541	*	*	no
Regression 2:									
Northeast	10.9	0.204	0.403	10.9	0.204	0.403	0.426	0.441 [a]	no
Midwest	9.8	0.197	0.383	9.8	0.197	0.383	0.426	0.380 [a]	no
South	13.1	0.182	0.382	13.1	0.182	0.382	0.426	0.106 [a]	no
West	12.6	0.208	0.431	12.6	0.208	0.431	0.486	0.428 [a]	no
Regression 3:									
Metropolitan	11.2	0.108	0.217	11.2	0.108	0.217	0.182	0.216	yes
Nonmetropolitan [b]	14.2	0.248	0.464	14.2	0.248	0.464	0.729	0.603	yes [c]

[a] No parameters exist for the regions, so the coefficients for national estimates (b=10,380) are used for each region.

[b] Percent poor differs from the P-60 reported value of 14.3 because the metro/nonmetro status for 340 observations (0.25% of the sample) is censored from the public-use data files.

[c] Nonmetropolitan matches with the adjustment of increasing nonmetropolitan (a,b) coefficients by a factor of 1.5.

Notes: Regressions are estimated in Stata with the *svyreg* command. For all regressions the constant is suppressed. The OLS standard errors account for the sample weights, but do not correct for clustering or stratification. The synthetic corrections account for weights, strata, and clusters. In the section labeled "Analysis of Sample Means", the SRS standard errors correct for weights but not the stratification or clustering. The synthetic corrections account for all three design features. The "Table A" standard errors are found by dividing the reported (one-digit) 90 percent confidence interval by 1.645, while the Equation (2) standard errors result from following the Census methodology for estimating standard errors of percentages. Match a,b categories is the same as in Table 1.