PROPERTIES OF VARIANCE ESTIMATORS FOR THE 1995 CENSUS TEST

Machell Kindred Town and Robert E. Fay, U. S. Bureau of the Census[1]
Machell Kindred Town, U. S. Bureau of the Census, Washington, DC 20233

KEY WORDS: *Census adjustment, imputation, jackknife, census undercount.*

Abstract The 1995 Census Test includes two fundamental changes in census design: sampling for nonresponse and integrated coverage measurement (ICM), both to be tested as precursors for 2000. After a determination of occupancy status by the Postal Service, housing units not responding to the mail census will be sampled and survey estimation approaches employed, in contrast to an attempt to follow up all nonresponses as in previous mail censuses. For ICM, a subsample will be drawn to estimate the residual undercoverage of the census, and estimates of the undercoverage will be integrated into the final count. The estimation incorporates aspects of both ratio estimation and imputation.

This paper evaluates, using data from past censuses and Monte Carlo simulation, variance estimators developed for the 1995 test and some potential alternatives.

1. INTRODUCTION

The 1995 Census Test includes two fundamental changes in census design: sampling for nonresponse and integrated coverage measurement (ICM), both to be tested as precursors for 2000. Three sites were chosen for test, partially because of their low mail response rate from the 1990 census: six parishes in Northwest Louisiana; Paterson, New Jersey; and Oakland, California. The 1995 Census Test provides a first opportunity to evaluate operational features of these new census measurement concepts. In addition, the test experimentally evaluates key assumptions with important design implications for the 2000 census.

Past censuses have represented an effort to count everyone directly. In previous mail censuses, there was an attempt to follow up all mail nonresponses in order to complete the enumeration. In the 1995 design of sampling for nonresponse, the Postal Service provides a determination of occupancy status. Blocks or units of nonresponding occupied addresses are then sampled for followup, and estimation methods used to estimate characteristics for occupied nonresponding housing units not in the nonresponse sample and the number of vacants in the Nonresponse Follow-up (NRFU). The estimation incorporates aspects of both ratio estimation and imputation.

In the 1990 census, the Post Enumeration Survey (PES) was used to estimate the net undercoverage, reflecting both omissions and erroneous enumerations. The results were not included in the final data products from the 1990 census. Plans for the 2000 census call for estimates of net undercoverage to be integrated into the census process and incorporated into the final count and all data products.

The focus of this paper is to evaluate the variance estimators developed for the 1995 Census Test and some potential alternatives, and to provide recommendations for the 2000 census. The research incorporates data from the 1990 census and PES, Monte Carlo simulations, and VPLX to estimate the variance estimators.

2. SAMPLE DESIGN AND ESTIMATION FOR THE 1995 CENSUS TEST

The design for the 1995 Census Test incorporates sampling and estimation for NRFU and ICM. In Oakland, the test evaluates two methods of NRFU sampling, block and unit sampling. The estimation based on the sampling of nonrespondents includes using the data from the housing unit sample, the block sample, and the ICM sample. The two methods of estimation, Census Plus estimation and dual system estimation (DSE), are evaluated in the ICM in Oakland and Paterson.

The evaluation of the 1995 Census Test compares the two methods of sampling for NRFU and the two methods of estimation to facilitate the decision on the 2000 census methodology. The comparison of the Census Plus and DSE focuses on the fact that they have different underlying assumptions. The Census Plus assumes that ICM finds the truth and the DSE assumes that the independent roster of ICM is not necessarily the truth but assumes instead that omissions from the census and from the ICM reinterview are statistically independent.

ICM Sample Design. The sample selection for ICM requires five steps. The first step is to combine blocks to form clusters of about 30 or more housing units. Blocks with more than 40 housing units (HUs) are not combined with other blocks and are not split or subsampled. A separate stratum of small blocks is formed from small blocks (0, 1, or 2 Hus) where each block is a cluster by itself. The block clusters are the ultimate sampling units. Next the block clusters are stratified into groups such that the clusters within each group are similar in terms of racial and ethnic composition, and size of the cluster. These groups are called ICM sampling strata. All small clusters make up a separate ICM sampling stratum regardless of racial/ethnic composition.

After forming ICM sampling strata, clusters within each stratum are further stratified into smaller groups such that these smaller groups are similar in terms of the percent of minority renters, the percent of single unit structure, the percent of non-mail return HUs, the tool kit methods (special programs designed to improve coverage) used in the cluster, and the household population. The groups are used to split the site into two panels. Half of each group is assigned to each of two panels using sampling to provide an even split of clusters with respect to selected demographic characteristics important for coverage measurement. Samples are drawn from both panels for ICM by the same procuedures.

The block clusters are then sorted by panel within each ICM sampling stratum and then a systematic sample is selected. The number of block clusters selected for ICM is 150 in Oakland, 100 in Paterson, and 75 in northwest Louisiana. For operational ease, these clusters were sometimes partitioned into two or more clusters.

After the ICM sample has been selected, the next step is to select the samples for NRFU in each panel from the clusters not selected for the ICM (Isaki 1995).

Nonresponse Follow-up Sample. Nonresponse follow-up is performed on 100 percent of the blocks selected for ICM; however, in non-ICM blocks, a supplemental sample was selected for the NRFU. (In 2000, far more blocks will be selected for NRFU than for ICM.)

In Oakland, a sample of blocks for the NRFU was selected in panel 1 and a sample of HUs was selected in panel 2. In the other two sites, only blocks were sampled for NRFU in both panels because of the lack of funds. The methodology for selecting blocks is similar in all three sites.

Block Sample. The first step in the selection of sample is to assign panel 1 blocks in ICM strata to NRFU analysis sampling strata based on their similarity in terms of 1990 census race/Hispanic origin concentration. The race and origin groups are defined as follows: Black; Hispanic origin; Asian and Pacific Islander; rest of the population (includes White, American Indian, Eskimo and Aleut, and other).

After forming sampling strata, blocks were sorted in each sampling stratum by the number of nonrespondents in descending order and then a systematic sample of blocks was selected. (Navarro *et. al.* 1995)

Unit Sample. A unit sample was selected in the Oakland site only. This sample is a representative of nonrespondent housing units in non-ICM blocks in panel 2. First, the number of nonrespondent units in each of the blocks is determined. If there are fewer than 4 nonrespondents in a block then one nonrespondent unit is selected. No unit is selected from a block with 0 nonrespondents. For blocks with 4 or more nonrespondents, the nonresponding HUs within panel 2 are sorted by ICM stratum, NRFU stratum, District Office, Address Register Area, Block, and identification number, then a systematic sample of nonresponding HUs is chosen.

Estimation. Using ICM and NRFU samples, estimates for entire site and for the groups defined by age/sex/Hispanic ethnicity/tenure within site are determined after NRFU and again after the ICM data are collected and incorporated into the census numbers.

Post NRFU Estimation The post NRFU census estimate is the sum of counts from self-response (mail returns, Be Counted forms, Reverse Computer Assisted Telephone Interviews) and estimates from NRFU sample.

The NRFU sample counts are estimated using a statistical estimation approach. A weight based on the unconditional chance of a unit being selected into either the NRFU or ICM samples is assigned to each unit (Ikeda and Singh 1995). This weight is the number of units a sample housing unit represents in the population. The weights are then adjusted through ratio estimation to total occupied nonmail housing units within 11 analytic strata poststratifying blocks by racial composition and response rate. The counts for a HU will be multiplied by the corresponding weight to get the total counts for all units represented by the sample HU. The sum of these counts over all sample HUs provides counts from NRFU sample.

ICM Estimation. There are two ICM estimation methods, Census Plus and Dual System (DSE) (Schindler and Singh 1995). The Census Plus is the resolved roster (R-Sample) for each housing unit in each ICM block cluster. The resolved rosters, generated from the field reconciliation, contains all persons who lived at the address on Census Day. The DSE is the independent roster (P-sample) collected during the independent part of the ICM interview is another independent list. The Census Plus estimation permits the use of the Census report to aid in determining the resolved roster; therefore the resolved roster is expected to provide a more accurate account of the households within the ICM block cluster. The DSE estimates persons that were missed on both the Census and the P-sample but the Census Plus does not estimate the persons missed on both the Census and R-sample.

The P-sample is matched to the E-sample which is the original census households in the ICM block clusters. The Census is then adjusted by the ratio of the number of persons in the P-Sample over the number of persons matched from the E-Sample.

The two estimators will be compared to make a decision for the census 2000 methodology.

3. VARIANCE ESTIMATION STRATEGY

The 1995 Test includes experimental designs, namely, the comparisons of block vs. unit sampling in Oakland and of DSE vs. Census Plus estimation. Consequently, it will be important to derive variance estimates appropriate for the experimental analyses. Variances are also required for the more usual data products from a census, and for these estimates the typical issues of variance estimation for finite populations are important. For example, from a finite population perspective, persons included in the census count after being counted from a mail-return questionnaire do not contribute variance to the estimated census total. We first consider variance estimation from the perspective of finite population sampling.

Variances for finite population estimates. The combination of NRFU and ICM poses some interesting challenges for variance estimation. The ICM sample is selected first and both the NRFU and ICM treatments are applied, then the remainder of the site is restratified and sampled in a different manner for the NRFU operations from the non-ICM blocks. In other words, the situation is analogous to double sampling in reverse. A smaller sample, s_1, with sampling fraction f_1, is first selected on which the most complete set of measurements - NRFU (X) and ICM (Y) - are carried out. The NRFU sample, s_2, with sampling fraction f_2, is then selected from the remaining blocks with nonresponding households and only one of the operations - NRFU (X) - completed. Data on mail nonresponse households, X, from both samples, s_1, and s_2, will be used in NRFU estimation. Estimates of Y from s_1 will be used to estimate residual undercoverage of the census, with results to be applied both directly to the ICM sample itself and estimated for the entire non-ICM sample.

In 2000, the ICM sampling fraction, f_1, will be almost negligible, and simplifications of the variance estimator in 2000 may take advantage of that fact. For the 1995 test, however, the ICM sampling fraction, f_1, is significant, and the Census data from ICM blocks will also constitute a substantial fraction of the total sample employed for NRFU estimation. We also note that systematic samples were employed in both stages of sampling. We will first discuss variance estimation for finite population characteristics as if both samples had been stratified simple random samples without replacement, and then return to discuss the effect of systematic sampling.

Because the ICM sample, s_1, was selected first, traditional approaches to variance estimation for stratified samples apply to estimates based on only the ICM sample. For example, a stratified jackknife incorporating the finite population corrections could produce variances for estimates based on s_1. Let $U - s_1$ denote the non-ICM blocks, which is the universe from which the stratified NRFU sample, s_2, is selected. Under standard assumptions, two unbiased estimates of a finite population total are possible. One uses the data from s_1 only:

$$\hat{X}_1 = \sum_{i \in s_1} f_1^{-1} x_i$$

while the second conditions on s_1:

$$\hat{X}_2 = \sum_{i \in s_1} x_i + \sum_{i \in s_2} f_2^{-1} x_i$$

The data from the two samples are combined so that each sample x_i is given weight f^{-1}, where

$$f = f_1 + (1 - f_1) f_2$$

that is,

$$\hat{X} = (1 - f_2)\left(\frac{f_1}{f}\right)\hat{X}_1 + \left(\frac{f_2}{f}\right)\hat{X}_2$$

We have

$$\begin{aligned} Cov(\hat{X}_1, \hat{X}_2) &= E(Cov(\hat{X}_1, \hat{X}_2 \mid s_1)) \\ &\quad + Cov(E(\hat{X}_1 \mid s_1), E(\hat{X}_2 \mid s_1)) \\ &= 0 \end{aligned}$$

since

$$Var(\hat{X}_1 \mid s_1) = 0$$

$$Var(E(\hat{X}_2 \mid s_1)) = 0$$

Thus, a jackknife estimate of variance may be formed from:

$$\begin{aligned} v_J(\hat{X}) &= (1 - f_1)(1 - f_2)^2 v_J(X_1^*) \\ &\quad + (1 - f_2) v_J(X_2^*) \end{aligned} \qquad (1)$$

where $v_J(X_1^*)$ represents the usual jackknife variance estimate for X_1^*, the portion of the estimated total $\hat{X}$ based on s_1, and where, analogously, $v_J(X_2^*)$ represents the usual jackknife variance estimate for X_2^*, the portion of the estimated total $\hat{X}$ based on s_2. In other words, both X_1^* and X_2^* are based on applying the unconditional weight, f^{-1}, to the sample data from s_1 and s_2, respectively.

Consequently, the variance estimation strategy employs two sets of replicates - one for the ICM sample and one for the NRFU sample - with coefficient multipliers in the variance estimator to reflect finite population effects and the weighting of the two samples.

One implementation of eq. (1) in VPLX incorporates the factors in (1) into finite population corrections. An alternative implementation is more advantageous for replicating X and Y jointly. Each X value from the ICM sample is represented twice - once with weight f_2 / f and once with weight $(1-f_2) / f$. The first weight is assigned to values X without Y and are treated as constants in the replication, as if self-representing. The second weight is assigned to joint observations of X and Y and employed in a stratified jackknife incorporating the usual finite population correction based on f_1.

Although the argument just stated was in terms of sampling without replacement, the sampling was in fact systematic. We investigated an extension to replication (Fay and Train 1995) of a variance estimator for systematic samples, studied previously by Wolter (1985, ch. 7). The original estimator is based on squaring differences of successive pairs of sample observations with respect to the original order of selection. Thus, if sample estimates neighboring in the systematic sort are far less variable from each other than the overall variation in the stratum, the variance estimator picks up the apparent effect of the systematic sort on the variance of the resulting estimates and adjusts the estimated variance accordingly.

Variances for Unit Sampling: In Oakland, one panel employs unit sampling in place of block sampling. For purposes of variance estimation of site-level characteristics, the same strategies may be employed to estimate the conditional variance for the NRFU unit sample, again taking into account the systematic sample.

Variances for Experimental Hypothesis: The assignment of the sample to panels and application of two separate forms of sampling represents an additional level of randomization imposed on the design. Assignment of blocks to panels is not a permanent feature of the finite population but rather the result of the experimental design. Consequently, it is important to analyze the data according to the underlying design. We propose to analyze the findings in two ways: using an overall srs of clusters (blocks), omitting finite population corrections, to assess the variance of the estimated treatment contrast:

$$v_J(\hat{X}) = (1-f_2)^2 v_J(X_1^*) + v_J(X_2^*) \qquad (2)$$

and through pairing census blocks, one each from the unit and block sample designs, and computing the variance of the differences from the pairings. For example, paired blocks could be treated as a single sample cluster in calculation of the stratified jackknife variance.

Similarly, we plan to take advantage of the ability of replication to capture covariance relationships when estimating the difference between census plus and DSE.

4. METHODOLOGY FOR THE MONTE CARLO STUDY

Matching of 1990 Census Files and PES Files. In order to build a universe to resemble the 1995 Census Test Paterson site, we extracted the Paterson 1990 Census data and statistically matched it to the 1990 PES data for the entire United States. The matching of the 90 Census and 90 PES files resulted in a 1995 Census Test-like data file for the test site. We matched the two data files at the block level. Only the urban blocks from the 90 PES file were eligible for the match. The statistically matched blocks from the 90 PES data were not restricted to the Paterson area because of the small PES sample of blocks in 90; therefore, the Census blocks were matched to similar blocks across the United States. The blocks were matched on five different variables: block size by persons, proportion of owner/renters, proportion of mail returns, proportion of blacks, and proportion of Hispanics. Blocks with more than 20 percent Hispanic and 20 percent Black were combined into a single set for matching purposes.

Five match runs were required to completely match all the 90 Census Paterson blocks with a 90 PES block. The following is the sequence of match runs that took place:

Match 1:	All five variables were used to match the two data files (block size, owner/renter, mail return, black/Hispanic).
Match 2:	All five variables were used but the proportion of black and Hispanic was reduced.
Match 3:	The mail return variable was dropped. The other four variables were used to match.
Match 4:	The owner/renter variable was dropped but the mail return variable along with the other variables were used for the match.
Match 5:	Both the owner/renter and mail return variables were dropped.

Paterson has many blocks with a high proportion of blacks and a high proportion of Hispanics. The 90 PES file have very few blocks with this occurrence; therefore, a small number of 90 PES blocks contributed to a large amount of the Paterson matched blocks (duplicates of PES blocks). This means that a single PES block could serve as a match for more than one Paterson Census block. After the five match runs, all of the blocks in Paterson matched to a 90 PES block.

Monte Carlo Design Except for minor details, we implement the sample design used in Paterson for the 1995 Census Test, both the ICM and NRFU components, although based exclusively on 1990 data. We studied two NRFU estimators: a simple estimate of total based on a single ratio estimation cell and ratio estimation for 11 analytic strata. For each selected sample, we examined estimators based on eq. (1) for the stratified jackknife, the systematic sample version, and half-sample replication where each sampling stratum was split into 2 half-samples. For ICM, we considered only census plus estimation using three general groups, Black, Hispanic, and Others.

5. RESULTS

Figures 1 and 2 indicate the percent bias for NRFU estimation. As Figure 1 suggests, all variance estimators experience difficulty estimating the actual variance from systematic sampling. In fact, the actual variance of estimates based on the systematic samples exceeds the variance based on a stratified random sample of the same size. Consequently, the jackknife behaves as well as variance estimators designed to take the systematic sample into account.

Figure 2 shows results for stratified random sampling without replacement. All three estimators are almost unbiased for ratio estimation based on a single cell, but only the jackknife does not experience a significant downward bias for estimates based on analytic strata. We observe that the denominator of the ratio estimator - the estimated number of nonmail households based on the sample blocks, had a c.v. on the order of 20% in many analytic strata.

We also noted that the c.v.'s of the estimated variance was quite high.

ICM estimates had over twice the variance of NRFU estimates. The results for ICM variance estimation followed patterns similar to NRFU estimation.

Future Work Similarly, we intend to also demonstrate the suitability of the variance estimators to be implemented for the experimental analyses. We will assume no treatment effects for the block vs. unit design and insure that the variance estimates are well calibrated under the experimental and sampling design.

We plan to repeat this study for the Oakland and Louisiana sites.

After completing this work for the given 1995 Census Test sample design, we hope to experiment with alternative sample designs and estimators.

These preliminary findings indicate that the variance approach described by (1) is appropriate. The findings also suggest that the block sample design puts some strain on the assumptions of design-based sample inference. The distribution of nonmail households by block is quite skewed, and we hypothesize that alteration of the block sampling design to switch to unit sampling in blocks with large numbers of nonresponding households would be quite advantageous. We plan to investigate alternatives soon.

[1] This paper reports the general results of research undertaken by Census Bureau staff. The views expressed are attributed to the authors and do not necessarily reflect those of the Census Bureau. We thank George Train for computational assistance, and Larry Bates, Courtney Ford, Alfredo Navarro, Rita Petroni, Raj Singh, Eric Schindler, and other Census Bureau staff for providing data and suggestions used in our analysis.

REFERENCES

Fay, R. E. and Train, G. (1995), "Aspects of Survey and Model-Based Postcensal Estimation of Income and Poverty Characteristics for States and Counties," to appear in the *Proceedings of the Government Statistics Section,* Alexandria, VA: American Statistical Association.

Ikeda, M. and Singh, R. (1995) "Computer Specifications for Site by Panel Level Post Nonresponse Follow-up Estimates for the 1995 Census Test (Revision 1)," 1995 Census Test Memorandum Series IS-7, Bureau of the Census.

Isaki, C. T. (1995) "Methodology for Selecting an Nonresponse Followup Sample Design for Three Sites," Bureau of the Census.

Navarro, A., Bates L., Scott, J., Isaki, C. (1995) "Computer Specifications for Sampling the Nonrespondents in the 1995 Census Test (Block and Unit Samples)," Decennial Statistical Studies Division 1995 Census Test Memorandum Series #A - 17, Bureau of the Census.

Schindler, E., Singh, R. (1995) "Computer Specifications for Integrated Coverage Measurement Estimation for the 1995 Census Test - Revision 1," Bureau of the Census.

Wolter, K. (1985), *Introduction to Variance Estimation*, New York: Springer-Verlag.

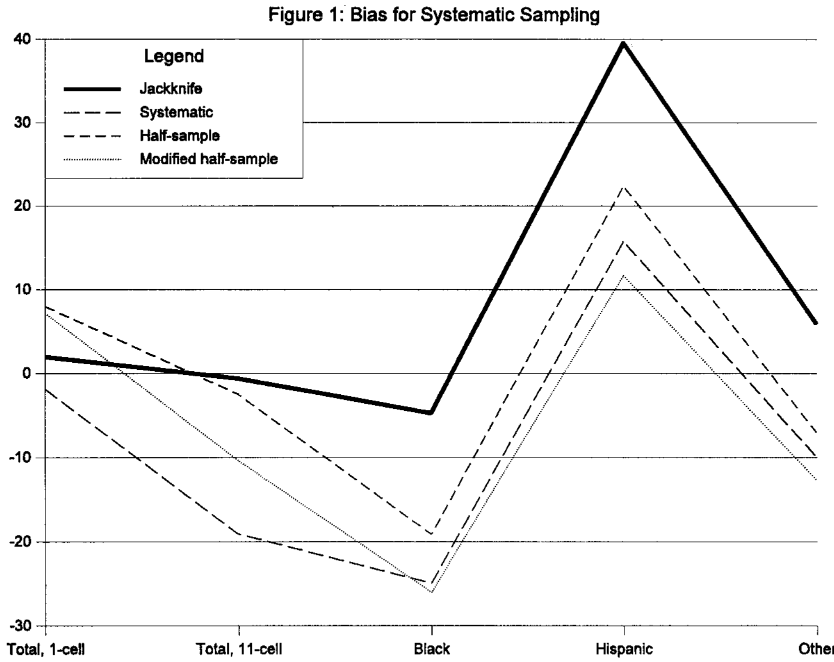

Figure 1. Percent bias in the estimated variance for systematic sampling, NRFU estimation.

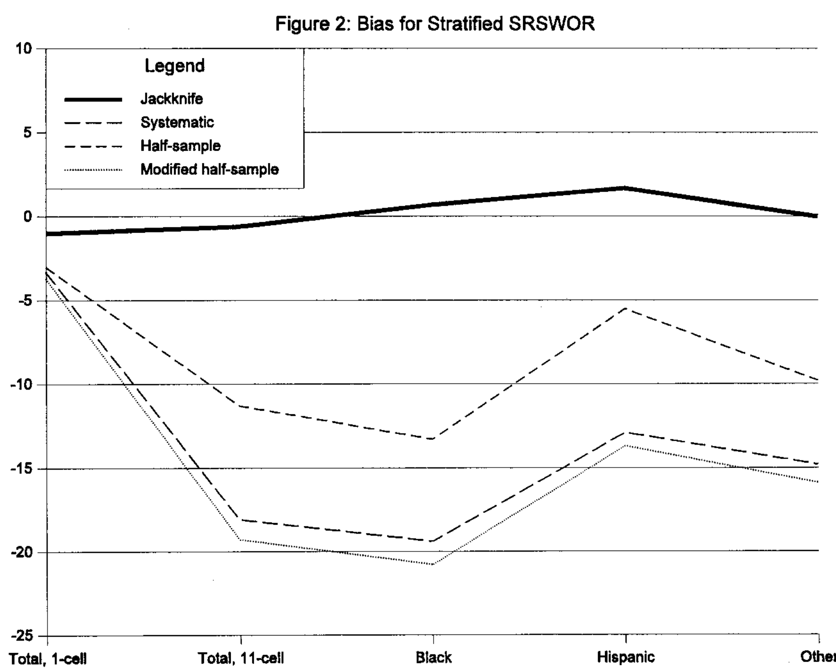

Figure 2. Percent bias in the estimated variance for stratified simple random sampling without replacement, NRFU estimation.