Survey Methodology, June 1995
Vol. 21, No. 1, pp. 55–62
Statistics Canada

# Between-State Heterogeneity of Undercount Rates and Surrogate Variables in the 1990 U.S. Census

JAY JONG-IK KIM, ALAN ZASLAVSKY and ROBERT BLODGETT[1]

ABSTRACT

As part of the decision on adjustment of the 1990 Decennial Census, the U.S. Census Bureau investigated possible heterogeneity of undercount rates between parts of different states falling in the same adjustment cell or poststratum. Five "surrogate variables" believed to be associated with undercount were analyzed using a large extract from the census and significant heterogeneity was found. Analysis of Post Enumeration Survey on undercount rates showed that more variance was explained by poststratification variables than by state, supporting the decision to use the poststratum as the adjustment cell. Significant interstate heterogeneity was found in 19 out of 99 poststratum groups (mainly in nonurban areas), but there was little if any evidence that the poststratified estimator was biased against particular states after aggregating across poststrata. Nonetheless, this issue should be addressed in future coverage evaluation studies.

KEY WORDS: Poststratification; Influence statistics; Linearization; Synthetic estimation.

## 1. INTRODUCTION

The Post Enumeration Survey (PES) of the 1990 Decennial Census of the United States was designed to produce coverage estimates for 1,392 poststrata. The nation was first divided into 116 domains, called poststratum groups (PSGs) according to geography, race/Spanish origin and tenure (owner *vs.* renter). With only 4 exceptions, all PSGs are defined within a census division, one of nine contiguous geographic areas each composed of several states. Each PSG was further divided into 12 age-by-sex groups, the poststrata. For example, roughly all Black renters in New York city constitute a PSG and all females, age 0-9, of this PSG make a poststratum (PS). Further details on the PES are in Hogan (1992,1993).

Small area undercount rates were calculated by synthetic estimation; the same adjustment factor was applied to persons from a given PS in all areas. This procedure is accurate under the "synthetic assumption" of homogeneity of undercount rate within a PS. The validity of the synthetic assumption has been hotly debated (Section 2). This paper reports on research conducted as a part of a PES evaluation project (the "P12 project") which investigated heterogeneity within poststrata. In particular, this research focused on the following question: can differences in coverage be identified between parts of a poststratum that fall into different states?

Under the homogeneity assumption, the rates are the same within a PS regardless of state. Thus, this assumption can be tested by comparing rates from state to state within a PS; this test focuses attention on the question of whether synthetic estimation is "unfair" to certain states. The unit of analysis is the intersection of a census block and a PS or PSG, called a block part (BP) for the analysis of the undercount rate data. A census block is a small area bounded by visible features such as streets, streams *etc.* and/or by political boundaries. In urban areas it roughly corresponds to a city block. In fact, most of our analyses are performed on PSGs, since the age-sex breakdown of the PSG did not vary much from state to state. Hence, the analysis focuses on whether BPs differ between states within PSG.

Two distinct analyses were performed. The distributions of five "surrogate variables" were investigated (Section 3), using a large (4.26%) extract from the census. The distribution of undercount was investigated using the much smaller PES data set (Section 4). For more detailed tables and documentation of the project, see Kim (1991).

## 2. LITERATURE REVIEW

Two key questions have been addressed in the literature on heterogeneity:

1. The empirical question: how much heterogeneity is there, and how can it be described?
2. The theoretical and policy question: what are the implications of heterogeneity for the accuracy of synthetic adjustments and the validity of assessments of these adjustments?

Heterogeneity may be identified and analyzed at many levels of aggregation. Perfect homogeneity of undercount rates for very small domains is numerically impossible,

[1] Jay Jong-Ik Kim, Statistical Research Division, U.S. Bureau of the Census, Suitland, MD 20233, U.S.A.; Alan Zaslavsky, Department of Statistics, Harvard University, Cambridge, MD 02138, U.S.A.; and Robert Blodgett, U.S. Food and Drug Administration, 200 C St., S.W., Washington, DC 20204, U.S.A.

because of discreteness of the true population and the census counts. Indeed, because census errors (omissions or erroneous enumerations) tend to be either independent of each other or positively associated (as when a household with several members is omitted, or when some local characteristic affects an entire block), we would anticipate at least binomial variability in observed undercount rates.

Hengartner and Speed (1993) analyzed 1990 PES data from two sites by fitting models in which the explanatory variables were block and "demoid" (a unit defined by the non-geographic poststratification variables, such as race, sex, age, and tenure). They found that the amount of variance explained by block was slightly greater than the amount explained by demoid; the number of blocks was not much greater than the number of demoids in their data set. In response, Schafer (1993) argued that an estimation scheme involving block effects would not be practical because it would require collecting data from every block.

Heterogeneity of undercount at any level may be defined as excess variability in observed undercount rates at that level over what would be expected as a consequence of variability at a lower level of aggregation. For example, confining our attention to a single poststratum, a set of blocks are heterogeneous if their undercount rates in that poststratum differ more than would be expected if households, including those counted, partially counted, and omitted in the census, had been randomly distributed across the blocks. Similarly, a group of states are heterogeneous (similarly controlling for poststratum) if they differ more than would be expected if blocks, including those with higher and lower undercounts, had been randomly distributed across the states. Several studies have attempted to measure heterogeneity in undercount rates and other census variables. Wachter and Freedman (1992) analyzed a large sample of census data (similar to that considered in Section 3). They estimated the excess variability between "superblocks" over that predicted by a binomial model with uniform rates, for four "artificial population" variables (multi-unit housing rate, nonmailback rate, allocations, and substitutions, described in Section 3). Compared to the greatest possible amount of heterogeneity (if each block were homogeneous), the "excess variability" ranged from around 20% (for multiunit housing) to 2% (for substitutions). Another study by Freedman and Wachter (1993) examined between-state heterogeneity using "artificial populations" based on the same variables and two others, and found substantial variability.

Alho, Mulry, Wurdeman and Kim (1993) used conditional logistic regression models to describe heterogeneity associated with measured covariates that were not captured in the poststratification. Their concern was primarily with reducing the bias of dual system estimates of population, rather than with more accurate small-area estimates.

A controversial topic in evaluation of the proposed adjustment of the 1990 census was the effect of heterogeneity on the accuracy of adjusted population counts obtained by synthetic estimation, and particularly on comparisons of the accuracy of adjusted and unadjusted counts. Wachter and Freedman (1992) argued that because the "synthetic assumption" of uniform coverage within poststrata is demonstrably false, aggregate measures of the accuracy of an adjusted census systematically underestimate error. Because nonuniformity of coverage affects the accuracy of an unadjusted census as well, however, the implications of this conclusion for the appropriateness of adjustment are not obvious.

In one of the earlier "surrogate variables" studies, Isaki, Schultz, Diffendal and Huang (1988) simulated the behavior of synthetic estimators on "artificial populations" which were transformations of the substitution (unit imputation) rate. They found that a synthetic estimator generally did better than "unadjusted" counts.

Schirm and Preston (1987) argued, using analytical calculations and simulation, that synthetic estimation makes estimates for small areas more accurate under plausible conditions, even if the synthetic assumption does not hold. Wolter and Causey (1991) investigated the performance of synthetic estimators and of a single ratio adjustment when the undercount rates are estimated with error, using undercount rates from the 1980 Post-Enumeration Program (PEP) and simulating various levels of sampling error; they estimated "break-even" coefficients of variation at which sampling error in the adjusted counts or proportions would make them less accurate than unadjusted counts or proportions. The conclusions of these articles were criticized by Freedman and Navidi (1992), who gave counterexamples of possible distributions of undercount for which adjustment by synthetic estimation would make population distribution less accurate.

Fay and Thompson (1993) simulated effects of heterogeneity on accuracy of synthetic estimates, using eight surrogate variables (including the five used in this study) and the same data set as analyzed in Section 3. They performed a loss function analysis as in Mulry and Spencer (1993) to compare the accuracy of simulated unadjusted counts to that of synthetically adjusted counts. They found that the effect of ignoring heterogeneity was to underestimate the gain in accuracy due to synthetic adjustment for five of eight variables, and to overestimate it for one variable (unemployment rate), while there was little difference for two other variables (poverty and migration rates).

## 3. ANALYSIS OF SURROGATE VARIABLES

In the analysis of census data, we selected variables which were available for the entire census and which, like undercount, were descriptive of or related to the census-taking process. The selected surrogates are the allocation rate, mail return rate, multiunit structure rate, mail universe rate (fraction of units receiving mail questionnaire) and substitution rate. The allocation rate is the fraction of households for which imputations were made for item nonresponse, and the substitution rate is the fraction of households which were imputed as a whole because it was determined that a unit was occupied but no interview could be obtained.

Table 1 shows correlations between each of these variables and undercount rate by PSG. These "ecological" correlations (Freedman, Pisani and Purvis 1978, pp. 141-142) of PSG averages differ from those which could be calculated from block-level data. The latter are smaller, possibly because of the noise introduced by random variability in the small populations in each block.

**Table 1**

Correlation Coefficients between the Surrogate Variable and Undercount Rate by PSG

| Variable | Correlation |
|---|---|
| Allocation Rate | .44 |
| Mail Return Rate | −.57 |
| Multiunit Structure Rate | .39 |
| Mail Universe Rate | .08 |
| Substitution Rate | .47 |

Applying a naive test which treats the PSGs as independent, each correlation is significant except that for mail universe rate, but the magnitudes of the correlations are not large. To some extent, furthermore, these variables are descriptive of conditions which tend to lead to higher omission rates (allocations due to poor completion of questionnaires, substitutions due to difficulty in obtaining interviews) or to lower omission rates (high mail return rates). On the other hand, difficult census-taking conditions can also lead to erroneous enumerations, so these effects on net undercount are not entirely clear-cut. We do not analyze these variables simply because we believe that they are distributed in exactly the same way as undercount. Rather we hope that by obtaining results on the distributions of a range of different census variables, we may gain some insight into the distribution of undercount.

For the analyses of the surrogate variables, a stratified cluster sample of 1990 Census data was extracted. This sample is composed of 204,394 blocks corresponding to 125,000 block clusters. A block part containing less than ten persons was combined with successive block parts (in order by block number) until a minimum count of ten persons was obtained. This operation was performed to obtain relatively stable rates for the surrogate variables which allows us to analyze the rates themselves.

Surrogate variables are analyzed by logistic regression. Two forms of logistic regression model were used. For the within-PSG analysis, the model for PSG $i$ is

$$log[P_{ij}/(1 - P_{ij})] = A + C_j$$

and for the within-division analysis,

$$log[P_{ij}/(1 - P_{ij})] = A + B_i + C_j,$$

where $P_{ij}$ is the rate for a surrogate variable in the $i$-th PSG and $j$-th state, $A$ is the intercept, $B_i$ is the $i$-th PSG effect and $C_j$ is the $j$-th state effect. The models used only the 99 PSGs astride two or more states. Models were built for surrogate variables in the 99 PSGs and in each of nine divisions. SAS PROC CATMOD estimated the parameters by maximum likelihood and provided Wald statistics for testing the significance of state effects.

The data were collected with a cluster sample rather than a simple random sample so the test statistics must be divided by a design effect. We estimate a design effect,

$$\hat{D}_{ij} = \frac{\sum_{k=1}^{K_{ij}} n_{ijk}(\hat{p}_{ijk} - \hat{p}_{ij})^2}{K_{ij}\,\hat{p}_{ij}(1 - \hat{p}_{ij})},$$

where $\hat{p}_{ijk}$ is the rate for the $i$-th PSG, $j$-th state and $k$-th combined BP; $n_{ijk}$ is the size of the combined BP; $K_{ij}$ is the sample number of combined BPs in the $i$-th PSG in the $j$-th state and $\hat{p}_{ij}$ is the estimated rate for the $i$-th PSG and $j$-th state. The fraction is the ratio of the observed between-block variance to that expected under binomial sampling.

The estimated design effect $\hat{D}_{ij}$ is a measure of within-state within-PSG heterogeneity. The more within-state heterogeneity there is, the greater the sampling variance of the state-level rate and the harder it is to detect a significant difference. The magnitude of the design effect thus affects the statistical power of the hypothesis tests.

The calculated design effect only approximates the required correction. First, $\hat{D}_{ij}$ sums over the combined BPs rather than individual BPs. Second, the sample is a stratified cluster sample, and most or all post-strata span several sampling strata. The formula is only strictly correct for an unstratified sample. Third, the correct effect involves off-diagonal (covariance) as well as on-diagonal (variance) terms, and the off-diagonal terms are omitted. To account for the above, the calculated design effects were multiplied by the judgmentally chosen factor, 1.25.

A design effect was calculated for each surrogate variable and PSG. It is small (around 2) in most PSGs for the allocation and substitution rate. The effect is slightly higher for mail return rate, but it tends to be large (as much as 20) for multiunit structure and mail universe rate, since these characteristics are usually fairly uniform within a block but vary greatly between blocks.

Table 2 summarizes the design-corrected tests for state effects within PSG.

**Table 2**

Number of PSGs with Significant ($\alpha = .05$) State Effect (Logistic Regression)

| Div. | No. Grp | Alloc | Mail Ret | Mult Str | Mail Unv | Sub |
|---|---|---|---|---|---|---|
| 1 | 5 | 5 | 5 | 5 | 1(1) | 3(4) |
| 2 | 12 | 11 | 11 | 12 | 7(10) | 12 |
| 3 | 16 | 15 | 16 | 16 | 3(3) | 12(12) |
| 4 | 8 | 8 | 8 | 7 | 5(6) | 5(8) |
| 5 | 10 | 10 | 9 | 10 | 4(4) | 7(8) |
| 6 | 15 | 15 | 13 | 15 | 5(7) | 15 |
| 7 | 9 | 8 | 9 | 9 | 4(4) | 8(8) |
| 8 | 7 | 7 | 7 | 7 | 2(3) | 6(6) |
| 9 | 17 | 15 | 14 | 14 | 5(5) | 6(12) |
| Sum | 99 | 94 | 92 | 95 | 36(43) | 74(84) |

The numbers in ( ) are the number of PSGs for which test statistics are available when they are less than the number of groups.

Nationally, for each surrogate variable the state effect is significant for at least 84% of the PSGs. (The total number of PSGs varies because when a PSG falls entirely within one state or when only one state has non-zero observations for a particular variable, the corresponding model cannot be fit). The results are summarized at the division level. (Divisions 1 through 9 are New England, Mid-Atlantic, South Atlantic, East South Central, West South Central, East North Central, West North Central, Mountain and Pacific Divisions.)

Table 3 shows the magnitude of state effects, expressed as $\chi^2$ values of test statistics adjusted for design effect, for three variables having relatively high correlation with the undercount rate. In the table, the $\chi^2$ values have from 1 to 8 degrees of freedom.

**Table 3**

Magnitude of State Effects with respect to Test Statistics

|  | Allocation Rate | Mail Return Rate | Substitution Rate |
|---|---|---|---|
| Minimum | 4.3 | 0.28 | 5.46 |
| 25%-ile | 27.5 | 102.83 | 49.80 |
| 50%-ile | 68.9 | 254.49 | 97.35 |
| 75%-ile | 140.3 | 644.05 | 260.88 |
| Maximum | 945.2 | 8,779.88 | 1,815.12 |

In division-level models with state and PSG effects, both the state and PSG effects were significant at the 1% level in every division and for every variable (excluding mail universe rate in two divisions where a test statistic could not be calculated).

## 4. ANALYSIS OF UNDERCOUNT RATE

The results described above for surrogate variables were obtained early in the census process, but they have limited relevance to homogeneity of undercount itself. After PES data were processed, direct analysis of the distribution of undercount became possible.

The data set for these analyses merged two data sets for the 12,124 PES sample blocks, one for the *E*-sample (Census follow-up) and the other for the *P*-sample (PES). There were 12,124 collection blocks, some of which were split up for tabulation, giving 12,964 tabulation blocks. More importantly, because some of the smaller blocks were combined in the sampling, there were 5,293 block clusters sampled. Correct enumerations and *E*-sample total counts are on the *E*-sample file. The *P*-sample file includes *P*-sample total counts and counts of matches (*P*-sample cases that were included in the Census).

### 4.1 Variance Explained by State and PSG

For each division, a two-way ANOVA was fitted to undercount rates for state parts. Table 4 shows the ratio of the sum of squares due to PSGs to that due to states within a division.

**Table 4**

Variance of Undercount Rate Explained by State and PSG

| Div. | No. of Groups | No. of States* | SS (Group) / SS (State) | MS (Group) / MS (State) |
|---|---|---|---|---|
| 1 | 5 | 6 | 4.51 | 5.64 |
| 2 | 12 | 3 | 4.88 | .89 |
| 3 | 16 | 9 | 12.69 | 6.77 |
| 4 | 8 | 4 | 8.73 | 3.74 |
| 5 | 10 | 4 | 8.17 | 2.72 |
| 6 | 15 | 5 | 7.67 | 2.19 |
| 7 | 9 | 7 | 2.78 | 2.09 |
| 8 | 7 | 8 | 1.31 | 1.53 |
| 9 | 17 | 5 | 40.28 | 10.07 |

* States include D.C.

The ratio is always greater than one and in Division 9 it is 40.28, showing much larger effects for PSG than for state. The mean square for group also exceeds the mean square for state in each division except Division 2. This supports the decision to use the PS rather than the state as the cell for undercount estimation and adjustment.

### 4.2 Tests for State Effects on Undercount Rates

Assuming the substitution rate (fraction of units imputed for nonresponse) is negligible, the adjustment factor ($\hat{R}$) for a domain is

$$\hat{R} = \frac{WCE/WE}{WM/WP},$$

and the undercount rate is

$$1 - 1/\hat{R},$$

where $WE$ and $WP$ are the estimated population sizes weighted up from the $E$ and $P$-sample, respectively. $WCE$ is the weighted number of correct enumerations and $WM$ is the weighted number of matches in the PES.

The statistic for the influence (see Appendix) of the $i$-th BP on the adjustment factor or undercount rate is

$$I_i = \hat{R}\left(\frac{WCE_i}{WCE} + \frac{WP_i}{WP} - \frac{WE_i}{WE} - \frac{WM_i}{WM}\right),$$

where $WCE_i$, $WP_i$, $WE_i$ and $WM_i$ are contributions from the $i$-th BP to the totals above.

A linear model was fitted to BP influence statistics to test for state effects. Under the null hypothesis, all the state parts in a PSG have the same undercount rate and the expected mean of the influence statistics for each state is 0 within each PSG. The influence statistics can be analyzed with one way ANOVA within a single PSG or two way ANOVA for all PSGs within a division.

Table 5 summarizes the tests for state effects on linearized statistics within each PSG.

**Table 5**

Analysis of Linearized Undercount at the PSG Level

| Division | Number of PSG | Number of PSG with $P < .05$ |
|---|---|---|
| 1 | 5 | 0 |
| 2 | 12 | 3 |
| 3 | 16 | 4 |
| 4 | 8 | 5 |
| 5 | 10 | 2 |
| 6 | 15 | 1 |
| 7 | 9 | 0 |
| 8 | 7 | 1 |
| 9 | 17 | 3 |
| Sum | 99 | 19 |

The tests reveal significant heterogeneity between states in 19 out of 99 groups at the 5% significance level. The magnitude of the estimated state effect ranges from a few percent up to 20%, but the standard errors of these estimates are very large.

Table 6 summarizes the results of these analysis by place type. Place types 0, 1, 2 and 3 are large central cities in a Primary Metropolitan Statistical Area (PMSA), place types 4, 5 and 6 are non-central cities in PMSA with large central cities and place types 7, 8 and 9 are other areas.

The significant results are concentrated in PSGs for small areas (place types 7, 8 and 9). Ten out of 32 such groups show significant interstate heterogeneity at the 5% level. This suggests that the poststratification can be improved in those areas.

**Table 6**

Summary of Analysis of Linearized Undercount by Place Type

| Place Type | Number of PSG | Number of PSG with $P < .05$ |
|---|---|---|
| 0 | 11 | 3 |
| 1 | 23 | 1 |
| 2 | 12 | 1 |
| 3 | 8 | 1 |
| 4 | 0 | 0 |
| 5 | 6 | 2 |
| 6 | 6 | 1 |
| 7 | 11 | 3 |
| 8 | 11 | 4 |
| 9 | 10 | 3 |

Table 7 shows the $F$-statistics and $p$-value for state effect for state × PSG models, once weighted by the size of domain and once without weights.

**Table 7**

State Effects by Division - Weighted and Unweighted Data

| Division | D.F. | Unweighted Models | | Weighted Models | |
|---|---|---|---|---|---|
| | | $F$ | $p$ | $F$ | $p$ |
| 1 | 5 | .57 | .72 | .40 | .85 |
| 2 | 2 | 4.64 | .01 | 1.72 | .18 |
| 3 | 8 | .43 | .91 | .65 | .74 |
| 4 | 3 | .64 | .59 | .66 | .58 |
| 5 | 3 | .66 | .58 | 1.37 | .25 |
| 6 | 4 | .60 | .66 | .24 | .92 |
| 7 | 6 | .39 | .88 | .22 | .97 |
| 8 | 7 | .62 | .74 | .76 | .62 |
| 9 | 4 | .77 | .54 | .48 | .75 |

The additive effect of state was significant in only one division ($p = .01$) in the unweighted state × PSG model; when data were weighted by size of domain, the smallest $p$-value for the state effect was .18. In both cases, the most significant effect was observed in Division 2, in which New Jersey appeared to have higher undercount rate, controlling for PSG, than New York. Note that the most undercounted area in New York (New York City) had its own poststrata. In eight out of ten PSGs for which New Jersey and New York could be compared, including nonurban areas, the estimated undercount for New Jersey was larger than that for New York. Elsewhere, because the state effects in different PSGs varied in magnitude and sometimes in sign, and because only within a minority of PSGs in any division were there significant state effects, there was not significant evidence that in the aggregate the poststratification was biased against certain states.

Table 8 shows point estimates of the state effects in linear models for undercount rate by state part in each division, with effects for state and poststratum group. (Effects are centered at zero by division.) In effect, these are estimates of interstate differences after correcting for effects explained by the PSG composition of the different states.

**Table 8**

Estimated State Effects on Undercount within Division (as percent)

| **Division 1** | | **Division 4** | | **Division 7** | |
|---|---|---|---|---|---|
| CT | −2.42 | AL | −2.90 | IA | −1.10 |
| ME | .74 | KY | 1.89 | KS | −0.50 |
| MA | −0.48 | MS | −0.02 | MN | −0.01 |
| NH | −0.14 | TN | 1.03 | MO | −0.66 |
| RI | 1.43 | | | NE | 1.76 |
| VT | 0.90 | | | ND | −0.07 |
| | | | | SD | 0.60 |
| **Division 2** | | **Division 5** | | **Division 8** | |
| NJ | 4.18 | AR | 1.44 | AZ | 2.70 |
| NY | −3.91 | LA | −0.71 | CO | 0.68 |
| PA | −0.26 | OK | 1.58 | ID | −2.24 |
| | | TX | −2.30 | MT | −1.61 |
| | | | | NV | −0.10 |
| | | | | NM | 3.35 |
| | | | | UT | 0.08 |
| | | | | WY | −2.84 |
| **Division 3** | | **Division 6** | | **Division 9** | |
| DE | −0.42 | IL | 0.86 | AK | −0.78 |
| DC | 2.82 | IN | 1.12 | CA | 1.02 |
| FL | −0.88 | MI | −0.73 | HI | −0.18 |
| GA | −1.43 | OH | −0.88 | OR | −0.26 |
| MD | −1.32 | WI | −0.38 | WA | 0.18 |
| NC | 0.53 | | | | |
| SC | 0.70 | | | | |
| VA | −0.11 | | | | |
| WV | 0.11 | | | | |

The root mean square in the analysis of variance for state within division, averaged across all divisions, is 1.72 percent. Recall that only in the unweighted Division 2 analysis were the differences between states significant, it must be emphasized that the estimates in Table 8 do not represent well-measured interstate differences. The fact that the estimated effects are substantial in magnitude but are still not statistically significant tells us that the power of these tests to find interstate differences, given the sample sizes of the PES, is not as great as might be desired.

Another approach to the power problem is to consider the effect of reducing the size of the census extract used in analysis of surrogate variables by a factor of 25, the ratio of the census extract to the PES sample sizes. If we divide by 25 each of the chi-square test statistics summarized in Table 3, then in only 27 out of 99 PSGs would interstate differences have been significant for allocation rate (compared to 94 out of 99 PSGs with the full sample). Similarly, significant differences would have been found for 53 out of 99 PSGs for mail return rate (compared to 92 out of 99 PSGs with the full sample), and for 14 out of 84 for substitution rate (compared to 74 out of 84). Substitution rates are comparable in magnitude to undercount rates; after our hypothetical reduction of sample size, we obtain similar numbers of significant tests for substitution and undercount rates. It is plausible that with a much larger sample we would have found many more significant interstate differences, although one can only speculate on whether they would have been large enough to be of substantive concern.

## 5. DISCUSSION

This paper evaluates interstate heterogeneity in undercount rate and other census variables in the 1990 Census.

The evaluation used 1990 Census data and 1990 PES data. When this research was first embarked upon, the PES data were unavailable and were not expected to become available for analysis before the scheduled completion date. Surrogate variables from the 1990 Census were tested for significant heterogeneity among states within PSG. At the PSG level, state effect was significant ($\alpha = .05$) for 84%-95% of its PSGs for the various surrogate variables.

ANOVA on linearized undercount based on the PES data at the PSG level showed significant ($\alpha = .05$) state effects for 19 out of 99 PSGs. The significant results were concentrated in the PSGs in non-PMSA areas. Ten out of 32 such PSGs had significant state effects. This suggests that the poststratification in the relatively nonurban areas was not as successful as in the more urbanized areas.

How can we explain the different findings of the two studies? The two data sets had very different sample sizes, *i.e.*, the Census data had 125,000 block clusters but the PES data had 5,293 block clusters. It is therefore not surprising that small differences between states on surrogate variables would be statistically significant although corresponding differences would not be demonstrable with respect to undercount rates.

Furthermore, the correlations between the undercount rate and the surrogate variables are low as shown in Table 1. Therefore, any generalization from surrogate variables to undercount rates is somewhat conjectural. Given the modest correlation between undercount rates and surrogate variables, we prefer to give greater weight to the analysis of the PES data.

We conclude from these data that there are no demonstrable differences in average undercount rate between states within each division, after adjusting for PSG effects. While there is weak evidence for a difference between

New Jersey and New York within the Mid-Atlantic division, this result must be downweighted in the context of the number of divisions (nine) for which the test was performed. We conclude that if adjustment of population counts had been carried out based on the 1990 PES, no state would have been able to show that the poststratification was manifestly unfair in that it underadjusted that state relative to what direct state estimates showed that it deserved.

As the review in Section 2 shows, there is no consensus on whether or not between-state heterogeneity in undercount rates within PSG which is of substantial magnitude, although not large enough to be accurately measured by PES, would systematically affect the gain in accuracy obtained by synthetic adjustment. Nonetheless, the differences between states that were identified in analysis of the PES, together with the ancillary evidence of the surrogate variable analyses, make it appear likely that heterogeneity between states will again be an issue in coverage measurement for the year 2000 census, especially for the larger states for which these coverage differences can be most accurately measured. Fay and Thompson (1993) argue that a coverage measurement sample for 2000 should be designed to support direct (rather than synthetic) estimates of undercount for all states, although a CNSTAT panel (CNSTAT 1994) warns that for some states this could impose a highly inefficient sample allocation. Research over the intervening years must address the development of a combination of sample design and estimation methods that will produce defensible estimates of population by state.

## ACKNOWLEDGEMENTS

The authors express their appreciation to the referees, an associate editor and editor for their comments which led to an improved version of this manuscript. This paper reports general results of research undertaken by the authors. The views expressed are attributable to the authors and do not necessarily reflect those of the Census Bureau or Harvard University. Zaslavsky's work was supported by Joint Statistical Agreements 90-23 and 91-31 and a contract between the Census Bureau and the National Opinion Research Center. The third author was with the Bureau of the Census when the research was in progress.

## APPENDIX

### Testing for Interstate Differences Using Linearized Statistics

A two-way ANOVA for adjustment factors in state parts yields an intuitively meaningful summary of the relative contributions of state and PSG effects to the variation in adjustment factors. Because the sampling unit of the PES is the block cluster rather than the state part, these models do not yield valid statistical tests of the significance of the state effects.

Consider a statistic whose sample estimate for a state or state part is a weighted mean of the sample estimates in each component block or BP. Significance of the state effects for this statistic within a PSG could be evaluated by one-way ANOVA with the included block parts as units (corresponding to PSUs), or aggregated across PSGs by two-way ANOVA for state and PSG effects.

The sample adjustment factor estimate $(WCE/WE)/(WM/WP)$ is a nonlinear function of sample counts. In small primary sampling units (PSUs) such as block parts this nonlinearity may be very noticeable, especially when the number of matches in a PSU is very small or zero so that the sample estimate of the adjustment factor is large or infinite. In this situation, if PSU sample estimates are treated as data, the additive assumptions of ANOVA are violated. Useful tests may be recovered, however, by using a linearized version of the statistic of interest.

Suppose that we are interested in a parameter $Z = f(X)$ where $X$ is a vector of population proportions in certain cells. Let $\bar{x}$, $x_i$ represent the corresponding sample proportions in the entire sample and in PSU $i$ respectively, so $\bar{x} = \sum N_i x_i / \sum N_i$ is a size-weighted average of block cell proportions. Let $f_1(X)$ be the gradient of $f$ at $X$. Then by Taylor linearization $f(\bar{x}) - f(X) \approx f_1(X)'(\bar{x} - X) = \sum N_i f_1(X)' x_i / \sum N_i - f_1(X)'X$, *i.e.*, we may treat the problem as one of inference regarding the quantities (pseudo-observations) $z_i = f_1(X)' x_i$. Because the approximate (linearized) influence of PSU $i$ on the estimate $f(\bar{x})$, that is, the difference between the estimate with and without PSU $i$ included, is $N_i f_1(X)'(x_i - \bar{x})$, we may describe this as a method based on influence statistics (Hampel *et al.* 1986) or the infinitesimal jackknife (Efron 1982, Chapter 6).

To derive a sensible variance model, suppose that we may regard PSU $i$ as sample (not necessarily SRS) from a superpopulation with cell proportions $X_i$. A simple model is then, for some covariance matrices $U_i$ and $V_i$,

superpopulation model:

$$E(X_i) = X, \quad \mathrm{Var}(X_i) = V_i$$

and

sampling model:

$$E(x_i \mid X_i) = X_i, \quad \mathrm{Var}(x_i \mid X_i) = U_i.$$

The sampling covariance $U_i$ will typically be proportional to $N_i^{-1}$. A plausible and mathematically convenient specification for $V_i$ is $V_i \; \alpha \; N_i^{-1}$ (*i.e.*, smaller PSUs more variable than larger ones), so $\mathrm{Var}\, z_i = \sigma^2/N_i$ for some constant $\sigma^2$. The corresponding linear model weight for PSU $i$ is $N_i$ so the model-based estimate of the mean agrees with the design-based estimate obtained by aggregating the cell counts if $N_i$ is a weighted size measure.

In the case of the adjustment factor $\hat{R} = (WCE/WE)/(WM/WP)$, the pseudo-observations are of the form $z_i = f_1(X)'(x_i - \bar{x}) =$

$$\hat{R}\left(\frac{WCE_i}{WCE} + \frac{WP_i}{WP} - \frac{WE_i}{WE} - \frac{WM_i}{WM}\right),$$

where $WCE_i$, $WP_i$, $WE_i$ and $WM_i$ are similar to the above for the $i$-th BP. We approximate the appropriate weight of a block part by $N_i = (WE_i + WP_i)/2$.

If the variance specifications of the model are inaccurate so there is some heteroscedasticity, or if the distribution is very long-tailed, then there will be a long-tailed distribution of residuals, making the tests at least slightly liberal. Some care must be taken to note the presence of outliers signaling this heteroscedasticity, for example, outlying blocks due to large-scale geocoding errors.

The assumption of approximately independent observations in ANOVA may be violated in two ways. First, the PSUs are not selected by SRS but rather by a geographical stratification somewhat finer than reflected in the post-stratification scheme. To the extent that this geographical stratification reduces the sampling variance of the state effect estimates, inferences under the independence model will be somewhat conservative. Second, there will be correlations between adjustment factors for different block parts from the same block (in multi-PSG models). These will tend to make inferences assuming independence somewhat liberal. On the balance, we regard the tests performed here as useful.

## REFERENCES

ALHO, J.M., MULRY, M.H., WURDEMAN, K., and KIM, J. (1991). Estimating heterogeneity in the probabilities of enumeration for dual-system estimation. *Journal of the American Statistical Association*, 88, 1130-1136.

BUREAU OF THE CENSUS (1990). Sample Selection Procedures for Performing Evaluation Study P12. STSD 1990 Coverage Studies and Evaluation Memorandum Series No. N-1, Memorandum from D. Bateman to L. Iskow and M. Lynch, October 3, 1990.

BUREAU OF THE CENSUS (1991). Request for Block Split Level Data for Performing PES Evaluation Project P12. STSD 1990 Coverage Studies and Evaluation Memorandum Series No. N-2, Memorandum from J. Thompson to A. Jackson, January 30, 1991.

COMMITTEE ON NATIONAL STATISTICS, PANEL TO EVALUATE ALTERNATIVE CENSUS METHODS (1994). *Counting People in the Information Age*. Washington D.C.: National Academy Press.

EFRON, B. (1982), *The Jackknife, the Bootstrap and Other Resampling Plans*. Philadelphia: Society for Industrial and Applied Mathematics (SIAM).

FAY, R.E., and THOMPSON, J.H. (1993). The 1990 Post Enumeration Survey Statistical Lessons, in Hindsight. *Proceedings of the 1993 Annual Research Conference*. U.S. Bureau of the Census, 71-91.

FREEDMAN, D.A., and NAVIDI, W.C. (1992). Should we have adjusted the U.S. Census of 1980? *Survey Methodology*, 18, 3-24.

FREEDMAN, D.A., PISANI, R., and PURVIS, R. (1978). *Statistics*. New York: Norton.

FREEDMAN, D.A., and WACHTER, K.W. (1993). Heterogeneity and Census Adjustment for the Inter-Censal Base. Technical Report No. 381, Department of Statistics, University of California at Berkeley.

HAMPEL, F.R., RONCHETTI, E.M., ROUSSEEUW, P.J., and STAHEL, W.A. (1986). *Robust Statistics: The Approach Based on Influence Functions*. New York: John Wiley and Sons.

HENGARTNER, N., and SPEED, T.P. (1993). Assessing between-block heterogeneity within the poststrata of the 1990 Post Enumeration Survey. *Journal of the American Statistical Association*, 88, 1119-1125.

HOGAN, H. (1992). The 1990 Post Enumeration Survey: An overview. *American Statistician*, 46, 261-269.

HOGAN, H. (1993). The 1990 Post Enumeration Survey: Operations and results. *Journal of the American Statistical Association*, 88, 1047-1057.

ISAKI, C.T., SCHULTZ, L.K., DIFFENDAL, G.J., and HUANG, E.T. (1988). On estimating census undercount in small areas. *Journal of Official Statistics*, 4, 95-112.

KIM, J. (1991). 1990 PES Evaluation Project P12: Evaluation of Synthetic Assumption. 1990 Coverage Studies and Evaluation Memorandum Series No. N-4, internal memorandum, U.S. Bureau of the Census.

MULRY, M.H., and SPENCER, B.D. (1993). Accuracy of the 1990 Census and undercount adjustment. *Journal of the American Statistical Association*, 88, 1080-1091.

SCHAFER, J.L. (1993). Comment on Hengartner, N and Speed, T.P.'s Assessing between-block heterogeneity within the poststrata of the 1990 Post Enumeration Survey. *Journal of the American Statistical Association*, 88, 1125-1127.

SCHIRM, A.L., and PRESTON, S.H. (1987). Census undercount adjustment and quality of geographic population distributions. *Journal of the American Statistical Association*, 82, 965-978.

WACHTER, K.W., and FREEDMAN, D.A. (1992). Measuring Local Homogeneity 1990 Census Data. Technical Report, Department of Statistics, University of California at Berkeley.

WOLTER, K.M., and CAUSEY, B.D. (1991). Evaluation of procedures for improving population estimates for small areas. *Journal of the American Statistical Association*, 86, 278-284.