Comment

Eugene P. Ericksen, Stephen E. Fienberg and Joseph B. Kadane

Our discussion of these three papers focuses primarily on those by Breiman and by Freedman and Wachter. Our observations are consistent with many of those made by Belin and Rolph in their paper, but this should surprise few people who are aware of both our previously stated positions on undercount adjustment and the role that two of us (Ericksen and Fienberg) had as expert witnesses in the recent New York City census adjustment litigation, described in part by Belin and Rolph.

1. BREIMAN ON BAD DATA

Until recently, many Census Bureau and other experts on census coverage equated the net undercount rate with the omission rate. For example, in a report describing the coverage of the 1970 census, the Bureau of the Census (1975) analyzed the consistency between the demographic estimate of undercount and the omission rate given by a postenumeration survey. They made no mention of the possible existence of erroneous enumerations, even though the bureau measured such errors as part of its first such survey in 1950.

In 1980, Bureau Director Vincent Barabba, in explaining his decision not to adjust the census gave as one of two reasons the fact that the net undercount was close to zero. Later analysis exposed the problem with his conclusion. First, the bureau reckoned that about 3 million undocumented aliens had been left out of the demographic estimate of the national population which was the basis of the conclusion that there was no undercount. Second, survey data collected in the 1980 Census Postenumeration Program indicated best estimates of 13 million omissions and 10 million counting errors, which are the sum of substitutions and erroneous enumerations. There appeared to be substantial variations in net undercount rates among places. The fact that the national estimates of omissions and counting errors were close was now seen as accidental.

In 1990, the situation was similar to that in 1980, but worse. The redefined question then asked by the Census Bureau was whether or not the observed geographic distributions of omissions and erroneous enumerations were real or were caused by errors in the data. In 1990, the estimated numbers of omissions (20 million) and counting errors (16 million) were much larger than they had been in 1980 (Bryant, 1993; Ericksen and DeFonso, 1993), and the net undercount (4 million) was slightly, but coincidentally, larger than it had been in 1980.

Given this background, the focus of Leo Breiman's paper seems misdirected. He concludes "The largest part of the original undercount estimate is due to bad data and processing error—80% on the national level." In Breiman's terms, he believes that the correct estimate of net undercount may be as low as 1 million. For this to be true, either the estimated number of omissions would have to be lowered from 20 to 17 million, the estimated number of counting errors would have to be increased from 16 to 19 million or there would have to be some combination of the two. Either way, there would be 30–40 million census errors to be accounted for, and if a decision was made not to adjust the census, one would simply have to hope that the distributions of these errors were so similar that between-area variations in net undercount rates would be minor. In our view, Breiman focused his time and energy on the wrong problem. Rather than trying to show how PES data problems inflated the national estimate of net undercount, he would have better spent his time showing how these errors might have skewed the estimated differentials between places.

Breiman's paper is based largely on the 1990 Post Enumeration Survey evaluation data, which came from three sources: (1) records of quality control procedures; (2) a repetition of matching procedures carried out for a sample of PES cases by more expert matchers at the Census Bureau; and (3) the Evaluation Followup Survey, in which a subsample of PES respondents were reinterviewed. This interviewing occurred in January 1991, fully nine months after census day and five to six months after the PES interviewing period. Using these evaluation data, Census Bureau statisticians had already assessed the quality of the PES data used for the undercount estimates. This evaluation was summarized by Mulry and Spencer (1993). In their best judgment, the national net undercount was *slightly* too high, but the differential undercount among places was substantially as the original PES had indicated.

Eugene P. Ericksen is Professor of Sociology and Statistics, Department of Sociology, Temple University, Philadelphia, Pennsylvania 19122. Stephen E. Fienberg and Joseph B. Kadane are Professors of Statistics and Social Sciences, Department of Statistics, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213.

TABLE 1

Weighted proportionate distribution showing how production PES E-sample classifications were reclassified in the evaluation study

	Original E-sample classification			
Evaluation study classification	Correct enumeration	Erroneous enumeration	Unresolved	Total
Correct enumeration	92.8	27.6	90.0	83.1
Erroneous enumeration	3.6	67.2	8.3	13.1
Unmatchable	0.01	—	—	0.01
Unresolved	3.6	5.2	1.6	3.8
Total	100.0	100.0	99.9	100.0
Sum of weights	28,005,586	4,990,529	774,766	33,770,880

Source: Table 35, U.S. Bureau of the Census, 1990 Coverage Studies and Evaluation Memorandum Series, #K-2, July 11, 1991.

Breiman feels that the Census Bureau underestimated the defects in the PES data.

Statisticians evaluating the quality of the adjustment to the 1990 census were generally aware of two shortcomings of the interview data used to evaluate the PES. Secretary of Commerce Mosbacher's Special Advisory Panel, a group whose members had sharply disparate views on adjustment, collectively wrote a letter to the secretary advocating great caution in the use of these data. Since the data were collected so long after census day, there would be uncertainty not just because of sampling but also due to nonsampling error. Just as there was substantial movement between census day and PES interviewing, many PES sample members moved between the day of their PES interview and the evaluation survey interviewing period. This was especially true of the hard-to-count members of the population.

Moreover, there was no direct way to measure correlation bias resulting from the fact that especially hard to count people are missed both by the census and the PES in excess of the numbers expected were the two independent. Breiman, in his evaluation, assumes that the correlation bias is zero. Mulry and Spencer, in accord with other Census Bureau statisticians, regard this assumption as unreasonable and used their best estimate of correlation bias. This dispute matters, because most of the errors cited by Breiman reduce the net undercount. Incorporating an estimate of correlation bias takes us the other way. We believe, along with Mulry and Spencer, that Breiman's assumption of zero correlation bias leads him to overstate his case. This overstatement is extended because he makes no allowance for the uncertainty caused by problems in collecting the evaluation survey data. No doubt those people missed either by the census or the PES were even harder to find in January 1991 during the evaluation survey interviewing period. We can only speculate as to why Breiman chose to believe an assumption that most knowledgeable statisticians have found untenable. In the original version of this paper, prepared for the 1992 New York City census trial, Breiman did present an argument for this position based on a fallacious manipulation of confidence limits across several different methods for estimating the impact of correlation bias (Bell, 1993). He has dropped the discussion but kept the erroneous conclusion.

Because Breiman studied the wrong problem, the national net undercount as opposed to the distribution of this undercount, we do not feel that his conclusions matter greatly. We prefer to rely upon the evaluation made by the Census Bureau, partly because they have studied the questions more thoroughly, but also because we believe them to be more objective. Many of Breiman's judgments appear to be strained, and his evaluations extreme. Here is an example.

In his Table 8, Breiman shows that when additional information was obtained from the Evaluation Followup interviews, 7.2 percent of cases originally classified as correct enumeration and 32.8 percent of cases originally classified as erroneous enumeration had their status changed as a result of the new information. That cases would change status as the result of new information does not surprise us, but we wish that he had presented a more complete analysis. As we show in Table 1, 13 percent of cases changed status; 6 percent went from unresolved to a resolved status or vice versa, while 7 percent went from erroneous to correct enumeration or vice versa. There was a fair amount of cancellation in these changes, and the correct enumeration percentage went from 82.9 to 83.1 percent. The erroneous enumeration percentage dropped from 14.8 to 13.1 percent while

TABLE 2
Estimated original dual system estimates and alternative correction estimates by evaluation strata

	Estimate			
	Original	[P16] corrected		
Evaluation poststratum	dual system	correlation bias	no correlation bias	Breiman
1. NE central-city minority	6.83	7.08	5.32	3.7
2. NE central-city nonminority	−0.75	−1.34	−1.34	−2.5
3. U.S. non-central-city minority	5.43	4.51	3.89	2.6
4. NE non-central-city nonminority	0.01	−1.03	−1.03	−1.6
5. South central-city minority	5.68	3.77	3.12	1.5
6. South central-city nonminority	1.94	1.14	1.13	0.3
7. South non-central-city nonminority	1.82	1.78	1.55	0.6
8. MW central-city minority	3.97	5.04	4.17	3.0
9. MW non-central-city nonminority	1.28	0.65	0.46	−0.5
10. MW non-central-city nonminority	0.39	−0.13	−0.13	−0.7
11. West central-city minority	6.14	6.38	5.72	3.9
12. West central-city nonminority	2.13	3.99	3.81	3.1
13. West non-central-city nonminority	1.84	0.68	0.68	−0.2
	Differences between evaluation poststrata			
Northeast (1–4)	6.82	8.11	6.35	5.3
South (5–7)	3.86	1.99	1.57	0.9
Midwest (8–10)	3.58	5.17	4.30	3.7
West (11–13)	4.30	5.70	5.04	4.1

Note: Evaluation poststratum 13 includes a poststratum of American Indian reservations.
Source: Original and [P16] corrected estimates are from U.S. Bureau of the Census, 1990 Coverage Studies and Evaluation Memorandum Series #R-6, July 11, 1991, and have been reproduced in Mulry and Spencer (1993). Breiman's estimates are from his Table 16.

the unresolved percentage increased from 2.3 to 3.8 percent. How much the final rate of erroneous enumeration would change depends on the proportion of unresolved cases imputed to be erroneous, but any way this turns out, the impact of these changes is minimal. Breiman's arguments here seem strained at best.

What Breiman has basically done is to repeat the Mulry and Spencer analysis, to argue that the estimate of zero correlation bias should be used and to add a few additional components of error. Aside from his assumption about correlation bias, his numerical results are not greatly different from those of the bureau, especially when we consider between-area differentials.

Mulry and Spencer (1993) used their total error model to adjust the net undercount estimates for 13 "evaluation poststrata," which were aggregates of the 1,392 poststrata defined by the PES. There were three evaluation poststrata in each of the four regions, one for minorities living in central cities and one for nonminorities living in central cities and one for nonminorities living outside central cities. The 13th stratum included all minorities living outside central cities in the nation. Of greatest practical interest is the comparison between central-city minority poststrata, where the net undercount was thought to be highest, and non-central-city nonminority poststrata, where the net undercount was thought to be lowest. In Table 2, we present the original estimates of net undercount computed by the Census Bureau; the same estimates corrected by Mulry and Spencer's total error model, with and without the Bureau's best estimate of correlation bias; and Breiman's final adjustments. At the bottom of the table, we present the differences between the central-city minority and non-central-city minority poststrata in each region. To evaluate the effect of assuming that the correlation bias is zero, compare the second and third columns. To evaluate the effect of Breiman's additional analysis, compare the third and fourth columns.

Looking at Table 2, we first see that for the original undercount estimates there were substantial differences between central-city minorities and non-central-city nonminorities in each region. When Mulry and Spencer adjusted these differences using the results of the total error model, the differences increased in three regions and decreased in just one, the South. Moving to column 3 and assuming zero correlation bias reduces all four differences. In three of the four regions, the differentials were within one percentage point of what they had been originally, but the change in the South was large. Finally, we observe that Breiman's additional adjustments reduced the differentials slightly. In the Northeast, Breiman's adjustment reduced the differential by 1.05%, and in the South, Midwest and West, the reduction was less than 1%. Comparing the original and Breiman's adjusted estimates (columns 1 and 4), Breiman reduced the net undercount differential by 1.5% in the Northeast and 3% in the South, but made scarcely any impact in the Midwest and West. Taking everything together, even if we were to believe Breiman's arguments and were to rely on his results rather than those of the Census Bureau, for the most part we find substantial differences in undercount rates between just those areas where we would expect to find them.

2. FREEDMAN AND WACHTER ON HETEROGENEITY

What do the Breiman and the Freedman–Wachter papers have in common? While both focus on the accuracy of adjustments to census data, they also adopt a shared implicit starting position, namely, that the census counts should be treated as if they are error free until we can show somehow that the use of adjusted counts is warranted. That the census is replete with errors and that the errors have differential impact on minorities never seems to be addressed or acknowledged by these authors.

The Freedman–Wachter paper (FW) addresses the issue of heterogeneity in the census adjusted counts when used for intercensal purposes. What they imply is that heterogeneity is bad and they tell us that "any comparison of error rates between the census and adjusted counts should take heterogeneity into account."

The argument of FW is that poststratum homogeneity is an assumption of adjustment. They find evidence of lack of poststratum homogeneity and seem to want to infer that adjustment is a bad idea. This latter does not follow, of course. Whether adjusted counts are closer to the truth than unadjusted counts is the essential issue and is only tangentially related to heterogeneity. Let us elaborate.

Heterogeneity among substrata is not inimical to adjustment. Kadane, Meyer and Tukey (1992) show that if substrata can be ordered by "catchability" in both the census and the PES simultaneously, the cross product ratio of the stratum is bounded *below* by an average (with respect to a certain probability weighting on the substrata) of substratum cross product ratios. In fact, the ordering is unnecessary: positive correlation (with respect to the same probability weighting) suffices. Thus if one accepts a cross product ratio of 1 in each substratum, as is often done in capture–recapture methods, and accepts positive correlation in catchability among substrata, the implication of substratum heterogeneity in catchability is that the use of a cross product ratio of 1 for adjustment moves in the right direction, but not far enough. A related paper by Darroch, Fienberg, Glonek and Junker (1993) proposes a model for individual-level heterogeneity which incorporates exactly this kind of dependence explored by Kadane, Meyer and Tukey (1992). Implemented in the context of triple system methods, their model when applied to census test data actually illustrates the extent to which heterogeneity in catchability might lead to underadjustment.

The conclusion of FW's analysis of surrogate variables is, as Belin and Rolph observe, at best overstated and they leave us confused about what this has to do with whether or not we might wish to adjust census counts for differential undercount. Fay and Thompson (1993), in the context of loss function analysis, use artificial substitutes for undercount rates as do FW. They conclude that "For most [proxy] variables, the loss function analyses is not seriously distorted in favor of adjustment, but in one case the loss function would overstate the advantages of adjustment." This is a far cry from FW's "almost anything can happen."

How are we to think about FW's proxy variables? Freedman and Wachter note that "The drawback is that the proxies may not behave like undercount rates, in terms of heterogeneity," but then they analyze them nonetheless. Perhaps they have in mind some model, of the form

$$\text{undercount rate} = f(\text{SUB}, \text{ALL}, \text{MULT}, \text{NM}, \text{MOB}, \text{POV}, \text{Error}).$$

The analysis of DIFF in their Table 9 suggests something of this form with f as a linear function with invented coefficients! What evidence do we have for believing such models and how much credence should we give to the analysis of the component parts?

Freedman and Wachter conclude that the 357 poststrata have too much residual variability on the proxy variables within state. This comes as no surprise to us. In its original analyses following the 1990 census, the bureau used 1,392 poststrata and then smoothed the resulting adjustment factors to remove variability. The later approach, which dropped the number of poststrata from 1,392 to 357, was sure to introduce greater heterogeneity and the "random" features in it are no longer kept under control through smoothing. As Fay and Thompson (1993) note: "Although [census analysts] eliminated smoothing from consideration in 1992, there may have been hidden costs to this decision. . . ."

The statistical literature clearly suggests that, even if FW were successful in showing substratum heterogeneity, by doing so they may have strengthened rather than weakened the case for the use of adjusted data.

Finally, we wonder what FW's analysis really has to say about the wisdom of using adjusted counts for various intercensal purposes. As we move further into the decade and away from the April 1, 1990, census date, there must be more and more error in census counts, both adjusted and unadjusted. Does there come a point when the cumulative errors due to the passage of time swamp the undercount problem? Or does the differential undercount between the white majority and various minority groups that we have observed for over 50 years in decennial census data only become worse?

Comment

Lars Lyberg and Sixten Lundström

The American census adjustment debate must represent the pinnacle of statistical methodological controversy. Usually, statistical discourse is conducted by laconic academics.who address technical issues of obscure merit to nonstatisticians. Indeed, it is an anomaly for the profession that an essentially technical issue, such as census adjustment, would attract such widespread and vocal attention.

Our comments should be prefaced with the fact the Swedish censuses are not affected by the type of undercoverage that characterizes the U.S. census undercount. Our approach to census taking is vastly different from that used in America. Sweden is known for its high-quality population registers and uses a register-based approach for the actual count. Since it is extremely difficult to function in Swedish society without a personal identity number (PIN) (and many of the benefits and amenities offered by Swedish society require a PIN), every legal resident is included in the population register. Any undercount is very small (a few hundred) and is linked to lags or delays in the reporting of vital statistics. These delays usually do not last more than 10–13 weeks; so, both in principle and practice, Sweden can conduct an accurate population census any week of the year. A word of caution, though, over the last few years, we have had an increasing problem with overcoverage due to immigrants who repatriate without notifying the authorities.

The United States, on the other hand, lacks Swedish-style population registers and bases its census on a master sample, that is, tracking everyone down by figuring out the number and location of dwellings and then ascertaining who and how many live in a given dwelling. This is obviously a daunting task when multiplied by an entire nation of geographically and ethnically diverse individuals.

When studying the articles by Freedman and Wachter, by Breiman, by Belin and Rolph and by numerous others dealing with the U.S. census adjustment, we have made a number of general observations. First, it should be kept in mind that statistics is the theory and practice of dealing with uncertainty. Second, surveys never produce "true" numbers. What surveys do produce are estimates and every source available should be used to make

Lars Lyberg is Head of Research and Development at Statistics Sweden and is Chief Editor of the Journal of Official Statistics. Sixten Lundström conducts work at the Statistical Research Unit at Statistics Sweden.