THE TREATMENT OF DIFFERENTIALWEIGHTS OF MOVERS IN THE PES

Eric Schindler and Richard Griffin,Bureau of the Census [1]

KEY WORDS: Monte Carlo, Taylor Series

INTRODUCTION

The Post Enumeration Survey or PES is a sample of approximately 5,400 clusters, either one or several blocks, containing 170,000 households. In the summer of 1990, these households were contacted to determine whether the occupants were correctly enumerated in the April 1990 Census. The recount of the persons in the blocks is called the P-sample. The verification of the April Census forms is referred to as the E-sample. Using the 1980 Census and the 1988 Dress Rehearsal as indicators, 3% to 5% of the E-sample, depending on age, sex, and race, was expected to be erroneously enumerated or to have inadequate data for verification. Similarly, 4% to 10% of the P-sample was expected to be missed by the Census and followup procedures. The analysis presented here was prepared during the fall of 1990 before PES data became available. A comparison of the assumptions with the actual PES data has been added.

The PES was selected from 105 geographically distinct sampling strata. Sampling rates vary from 1 in 70 to 1 in 2,700. Sampling rates for strata with more than 40% Blacks and/or Hispanics are about twice as high as for the non-minority strata in the same geographic areas. 116 basic estimation poststrata were formed by Census division, community type, race/origin, and tenure. Usually, three race/origin groups are used: Black, non-Black Hispanic, and Other. An example of a basic poststratum would be for Black renters in Type II MSA central cities in the South Atlantic region. Poststrata for Blacks and/or Hispanics often cover the Black and/or Hispanic population in several sampling strata. Each basic poststratum is divided into twelve final poststrata by sex and age to form 1392 estimation poststrata. A more complete description of the PES sample design is given in Woltman et al [1988].

Woltman et al assumed that overcount and undercount rates would be the same within each poststratum regardless of sampling stratum. However, the often wealthier minority persons in suburban areas of the Middle Atlantic, East North Central and Pacific Divisions with weights over 2,000 were missed only 11% of the time while minorities were missed 14.9% of the time overall.

People move. Approximately 8% of the P-sample moved to their PES addresses after Census Day. 80% of these moved within the same district office which may cover parts of several sampling strata. The original empirical study assumed a mover rate of 10% and that moving between sampling strata was uniform. This helped to insure an overestimate of the effect of movers on the estimates.

[1] This paper reports the general results of research undertaken by Census Bureau staff. The views expressed are attributable to the authors and do not necessarily reflect those of the Census Bureau.

Movers create additional problems. People in the process of moving may be unsure of which address to report at, may be too busy to report, may lose the form in the confusion of packing, and are more likely to be missed by normal Field Followup attempts. For the original study it was assumed that 8% of minority persons who did not move, but that 10% of moving minority persons, would be missed. Miss rates of 4% and 6% were assumed for non-minority nonmovers and movers.

If 10% of the 50,000 minority households in the PES moved between Census Day and the PES contact, 5,000 households moved. If10% were missed, and if all misses are clustered in households then there were about 500 missed mover households with about 1250 missed persons. About 1.5% of minority persons live in sampling strata with weights over 2000, so about 20 of these 1250 missed movers were sampled with these high weights. For estimation, these 20 persons willbe scattered across the 500 minority poststrata which have between 30,000 and 120,000 persons. Thus, a given minority poststratum has about a 4% (20/500) chance of having a missed mover who moved to a high weight sampling stratum. The minority poststrata have only about 6000 (10% of 60,000) movers of whom only 600 (10% of 6000) were missed. 2,000 missed movers will appreciably distort the estimates of misses (about 5,000 total) for these 20 poststrata.

Table 1 shows P-sample and E-sample data for Non-Hispanic Whites and all minorities. Miss rates were much higher than the rates based on previous data. It was also necessary to impute the probability of a match for about 2% of the sample. The miss rate and the imputation rate for those who moved from one district office to another were both much higher than for nonmovers or for movers within a district office.

Table 1. Actual PES Miss Rates and Erroneous Enumeration Rates

	Sample Size	Miss Rate (%)
P-Sample	377381	9.3
N-Minor	251212	6.5
N-Mover	231172	5.0
Movers	20040	22.9
SameDO	14123	19.4
New Do	5917	31.2
Minority	126169	14.9
N-Mover	116625	12.9
Movers	9544	39.1
SameDO	7380	34.0
New Do	2164	56.3
	Sample Size	E.E. Rate (%)
E-Sample	392587	5.5
N-Minor	259453	4.3
Minority	133134	7.9

MATHEMATICAL DEVELOPMENT

Define:

d to be the poststratum for which estimates are to be made.

h to be a sampling stratum which is within the geographic range of poststratum d.

i to be any sampling stratum in which a PES sample member who moved from h is found. For nonmovers, i = h.

P_{mdhi} to be the estimated miss rate for persons counted in poststratum d who lived in sampling stratum h on Census day but moved to sampling stratum i, which has weights w(i), before the PES.

P_{edh} to be the estimated erroneous enumeration/unmatchable rate for persons counted in poststratum d who lived in sampling stratum h on Census day.

CEN_{dh} to be the unadjusted Census count for poststratum d from sampling stratum h.

N_{dhi} to be the number of sample persons in sampling stratum i who lived in sampling stratum h on Census day and who are counted in poststratum d.

$$N_d=\sum_h N_{dh}=\sum_h \frac{CEN_{dh}*(1-P_{edh})}{(1-P_{mdh})}$$

to be the dual system estimator for poststratum d. It is the sum of the direct estimates for each poststratum. In fact, a combined dual system estimator is used where P_{ed} and P_{md} are calculated for each poststratum. One of the variations to Example 1 showed little difference between the two.

Since there are movers, the estimate of P_{mdh} is a weighted average given by:

$$P_{mdh}=\frac{\sum_i P_{mdhi}*n_{dhi}*w_i}{\sum_i n_{dhi}*w_i}$$

The weights of movers can be capped at some level, w' to control variance. In the literature, weight capping is usually accompanied by an associated increase in the uncapped weights by an appropriate factor to maintain totals. (See Potter, 1990, or many other papers.) For this model of the dual system estimator, the factor would be applied to both the numerator and denominator of P_{mdh}, so it can be omitted. For the combined dual system estimate, the move rates are low enough that the factors are close enough to 1 to be ignorable.

Let $w_i' = \min \{ w_i, w' \}$. Then an estimate of the revised miss rate is:

$$P'_{mdh}=\frac{\sum_i P_{mdhi}*n_{dhi}*w'_i}{\sum_i n_{dhi}*w'_i}$$

The revised dual system estimator is:

$$N'_{dh}=\frac{CEN_{dh}*(1-P_{edh})}{1-P'_{mdh}}$$

The difference between the estimates will be called the BIAS and is given by:

$$BIAS = N'_{dh}-N_{dh} = N_{dh}*\frac{P'_{mdh}-P_{mdh}}{1-P'_{mdh}}$$

Estimates of the expected variance are obtained using standard Taylor series methods taking the partial derivatives of N_{dh}' by P_{edh} and each P_{mdhi}.

$$VAR(N'_{dh}) = \frac{{N'_{dh}}^2*P_{edh}}{(1-P_{edh})*n_{dhh}/DEFF_{edh}} +$$

$$\frac{{N'_{dh}}^2}{(1-P'_{mdh})^2}*$$

$$\sum_i \frac{(n_{dhi}*w'_i)^2*\frac{P_{mdhi}*(1-P_{mdh}}{n_{dhi}/DEFF_{mdh}}}{(\sum_j n_{dhj}*w'_j)^2}$$

For unbiased estimates, no weights have been changed, the w' above can be replaced by the original w.

EXAMPLES

For the examples, five "pseudostrata", based on ranges of the sampling weights in the PES design, were formed for movers. All movers were assumed to have been sampled from one of these pseudostrata. The average weights were 283, 789, 1391, 1635, and 2337. Except for one variation of Example 1, it is assumed that 10% of the population of any poststratum consisted of outmovers, and, conversely, 10% of the sample in any sampling stratum consisted of inmovers. Design effects of 1.6 for erroneous enumerations and 3.5 for missing persons for minorities, and 1.1 for both erroneous enumerations and missing persons for non-minorities are taken from Woltman et al (1988).

Example 1 :

For a minority poststratum with a Census count of 60,000, assume that half are in sampling stratum 1 where 300 are sampled with weight 100, and that half are in sampling stratum 2 where 50 are sampled with weight 600. A 96% correct enumeration rate leaves 28,800 correct enumerations in each sampling stratum. 10% or 2880 moved to other sampling strata where some are sampled and returned to the correct poststratum for estimation. Using a 10% miss rate for movers, there are 288 missed movers per sampling stratum. The average PES weight is 600, so there should be about half a missed mover in the P-sample per poststratum. However, a missed mover with a weight of 2000 would add 2000 to the poststratum estimates. Weight capping reduces the effects induced by a single mover.

Table 2. shows dual system estimates, root mean square errors, and the BIAS for several weight caps for movers. 276 empirically produced the smallest RMSE.

Table 2. DSEs and RMSEs for several Weight Caps for a Minority Poststratum (assumed rates)

CAP	DSE	RMSE	BIAS
---	62748	3136	0
600	62711	2952	-37
276	62683	2925	-64
100	62637	2952	-111
0	62609	2992	-139

Table 3 shows the same poststratum using actual PES rates. A 7.5% mover rate, a 7.9% erroneous enumeration rate, and miss rates of 12.95% for nonmovers and 39.1% for nonmovers are used. The dual system estimates, mean square errors and biases are all considerably higher. The optimal weight cap resulted in a 12% reduction of the RMSE, but the cost was a bias equal to 17% of the adjustment and almost twice as large as the reduction in the RMSE.

Table 3. DSEs and RMSEs for several Weight Caps for a Minority Poststratum (actual rates)

CAP	DSE	RMSE	BIAS
---	65491	4728	0
600	64964	4232	-559
288	64561	4159	-930
100	63847	4285	-1644
0	63444	4456	-2047

Three additional variations of the above example were run with the assumed miss and erroneous enumeration rates to test the stability of the analysis. Even the most extreme variation produced only minor changes in the results.

Example 2 :

Assume a non-minority poststratum with a Census count of 600,000. Assume that 300,000 are in sampling stratum 1 where 750 are sampled with weight 400, and that 300,000 are in sampling stratum 2 where 300 are sampled with weight 1000. Assume an erroneous enumeration rate of about 3%, leaving 291,000 matches between the Census and the E-sample in each sampling stratum. Again assume that 10% of the matched sample persons in these sampling strata moved to other sampling strata (29,100 movers and 261,900 non-movers) where they are sampled and moved back to the correct poststratum for estimation. Assume a miss rate of 4% for the nonmovers and 6% for the movers in both strata. Assume that the movers are uniformly distributed across the five sampling pseudo strata. Because of the greater assumed stability and the larger sample sizes, only about a 2% reduction in the RMSE can be achieved with a bias of 380 out of 600,000. This example was repeated using the actual miss rates of 22.9% for movers and 5.0% for nonmovers, and the actual erroneous enumeration rate of 4.3%. The possible savings remain less than 3% of the rmse, but the bias necessary to achieve the savings is three times larger than for the assumed rates.

DATA SIMULATIONS

There was concern that the small number of movers missed by the Census in each poststratum might lead to a very non-normal distribution of population estimates. For example, a family with two young children which had moved away from a small poststratum to a sampling stratum with weights of 2,500. The two children would add 5,000 to the totals for the poststratum. If the family was missed, they would also add 5,000 to the number of missed persons in the poststratum. This could double the number of estimated misses. In order to investigate the distribution of estimates given the large weights of some movers, the situation was simulated.

One thousand simulations were made. The estimates for the unbiased and optimal cases for the smallest and largest five simulations are shown in Table 4. Capping the weights for the very large estimates causes larger adjustments than for the very small estimates.

Table 4. Simulations with Several Weight Caps for 5 Largest and 5 Smallest Unbiased Simulations

OBS	NO CAP	CAP = 272
1	56848	56312
2	57455	57890
3	57542	57715
4	57704	58288
5	58005	58197
996	67858	64976
997	67863	65925
998	68569	63553
999	68658	64680
1000	69030	66660

Variances were calculated using the standard formula. Because no design effects were used in these simulations the results are not comparable with those obtained above, but they are comparable with assuming design effects of 1.00 in the examples. The savings in RMSE attainable are of the same order of magnitude, i.e. 5%.

The distribution of the dual system estimates for the unbiased weights shows a slightly larger than normal tail on the high end, but a normal tail on the more critical (no one complains about being overcounted) low end. Using the optimal weight reduction scheme, or the other weight reduction schemes tried, generally reduced the simulated estimates which were very high, but did little for the very low estimates. It appears that low estimates of poststratum population would be no more frequent and no more serious with unbiased weights than with capped weights.

The changes in the dual system estimates between the optimal weight cap, N_{dh}', and the unbiased simulations, N_{dh}, were also examined. Using the optimal weight cap produced 564 increases and 436 decreases. For 176, or 40%, of the decreases, the unbiased simulated estimates are already smaller than the theoretical unbiased estimate of 62748. This could be as politically discomforting as the slightly larger sampling errors for the unbiased simulations.

CONCLUSIONS AND RECOMMENDATIONS

In conclusion, it is possible to obtain some savings in the mean square errors of the dual system estimate by capping the high weights of persons who have moved to areas with low sampling rates. However, based on the assumptions available in the Fall of 1990 when a decision had to be made, except for one improbable variation to Example 1, the theoretical savings were 7% or less of the root mean square error. Most of the improvement came from reducing the very high dual system estimates, not from increasing the low estimates. Also, capping weights reduced the unbiased estimates by more than 2% for about 10% of the simulations. The small gains did not justify the introduction of an additional source of bias into the dual system estimate. It was decided that the weights of movers would not be capped. Repeating the analysis showed that, because the actual miss rates for movers were considerably higher than those assumed, a recommendation to cap the weights would have been inappropriate because of the large biases.

REFERENCES

Woltman, H., Alberti, N., and Moriarity, C. (1988), "Sample Design for the 1990 Census Post Enumeration Survey," Proceedings of the Section for Survey Research Methods, 1988 meeting of the American Statistical Association.

Potter, F.J. (1990), "A Study of Procedures to Identify and Trim Extreme Sampling Weights," Proceedings of the Section for Survey Research Methods, 1990 meeting of the American Statistical Association.