Causes and Possible Remedies for Sampling Weight Variation in the Census 2000 Integrated Coverage Measurement Survey

Robert D. Sands and David E. McGrath
Bureau of the Census

Keywords: sample weights, weight variation, weight trimming and shrinkage, mean square error, dual system estimation

Introduction

The Census 2000 Integrated Coverage Measurement (ICM) Survey will be used to provide census totals designed to correct the undercount, especially a differential undercount among racial, ethnic, and socioeconomic groups, that has been observed in every decennial census since 1940. The ICM survey will be designed to produce direct estimates of total population for each of the fifty states and will have a sample size of 750,000 housing units. This paper will present results of research on the causes and proposed remedies for sampling weight variation in the Census 2000 ICM.

Theory of Weight Variation

Kalton (1983) notes that weights are used to compensate for the unequal probability of selection of sample units. Weights are also used in stratification after selection (poststratification) and in adjusting for nonresponse and noncoverage. A weighted sample mean estimator is actually a ratio estimator:

$$\bar{y}_w = \frac{\Sigma w_i y_i}{\Sigma w_i} \qquad (1)$$

where w_i is, in general, equal to the inverse of the probability of selection of sample element i. If the weights, w_i, are constant across the sample elements i, then the formula (1) reduces to the sample mean formula:

$$\bar{y}_w = \frac{\Sigma y_i}{n} = \bar{y}$$

The variance of (1) can be calculated using the Taylor linearization approach as follows :

$$var(\bar{y}_w) \doteq \frac{[var(u) + \bar{y}_w^2 var(w) - 2\bar{y}_w cov(u,w)]}{w^2} \qquad (2)$$

where:

$$u_i = w_i y_i,\ u = \Sigma u_i,\ w = \Sigma w_i$$

If it is assumed that the sum of the weights, $\sum w_i = N$, is constant across all samples in the design, then formula (2) becomes the following:

$$var(\bar{y}_w) = var(u)$$

In addition, if equal probability sampling is used, the weights, w_i, are constant across the sample elements i and formula (2) reduces to the variance of the sample mean:

$$var(\bar{y}_w) = var(\bar{y})$$

Moreover, if one applies the weighted mean formula to the stratified sample mean, the following formula is attained:

$$\bar{y}_w = \frac{\Sigma_h w_h \Sigma_i y_{hi}}{\Sigma_h w_h \Sigma_i 1} = \frac{\Sigma_h N_h \bar{y}_h}{\Sigma_h N_h} = \Sigma_h W_h \bar{y}_h = \bar{y}_{st} \qquad (3)$$

where the weight, $w_h = N_h/n_h$, is constant for all elements within a stratum h but varies for elements in different strata assuming disproportionate allocation of the sample to the strata. If we further assume that the population variance is each stratum, $S_h^2 = S^2$, is constant, then it can be shown that a sample employing weight variation as in (3) will have a higher variance than an equal probability sample of the same size.

Consequently, especially in sample designs employing

wide weight variation, methods to alleviate the problem have been investigated (Potter, 1988). In the present study, two methods will be evaluated using empirical data from the 1995 Test Census ICM. First, however, an illustration of the reasons for weight variation ib the ICM program will be given.

Sources of Weight Variation in the Census 2000 ICM

There are four potential sources of sampling weight variation in the ICM (Schindler, 1998). First, estimates calculated from person-level poststrata can have variable weights because persons in a given poststratum are divided among different block cluster-level sampling strata. These sampling strata may have disparate sampling weights. For the Dress Rehearsal (DR) Census, however, it has been decided to use proportional allocation of the sample to each sampling stratum. Proportional allocation will most likely also be used in Census 2000.

A second source of sampling weight variation is the result of the disproportionately low sampling rate of census blocks in the 'small' substratum. Furthermore, if blocks that are thought to be small (0, 1, or 2 housing units) are actually not small (10 or more housing units) a particular block can account for an abnormally large proportion of the weight of a particular poststratum group residing in the block.

A third source of weight variation stems from the sub-sampling of large block clusters. One segment of approximately 30 housing units is selected from the block clusters containing more than 80 housing units. The variable weight from the sub-sampling when multiplied by the first-stage sampling rate results in disparate weights. In the DR ICM, large block clusters were sampled at a higher rate in the first stage to account for the subsequent sub-sampling. This resulted in sampling weights for large block clusters very close to the weights for the remainder of block clusters. This plan, however, resulted in an overly burdensome listing work load which required a further sub-sampling of large blocks to take place between the first stage sampling and the sub-sampling. This second sub-sampling resulted in sampling weight variation from large blocks.

The final source of weight variation in the ICM is the result of differential nonresponse. In the ICM, weighting for nonresponse is done at the block cluster level resulting in weight variation within poststrata groups.

Possible Remedies for Weight Variation in the Census 2000 ICM Survey

Two approaches to reducing weight variation in the ICM Dual System Estimates of census undercoverage were investigated[1]. These approaches are (i) observing the effect of various levels of weight trimming on mean square error (MSE) of the estimates (Potter, 1988; Cox and McGrath, 1981) and (ii) computing optimal shrinkage weights as a function of within and between stratum variance and the unbiased stratum sampling weights (Stokes, 1989).

Estimated Mean Square Error (MSE) Trimming

One procedure to lessen the variance due to weight variation is to trim the weights and distribute the excess to other strata while maintaining the original sum of the weights. Potter (1988) refers to this technique as Estimated Mean Square Error (MSE) Trimming. It is well known that when sampling weights deviate from the inverse of the probability of selection bias in the estimates can occur (Cochran, 1977). On the other hand, trimming the larger weights can result in lower variance. The objective is then to minimize the MSE (the sum of the variance and the square of the bias) by trimming the weights to the point where the decrease in sampling variance is not yet eclipsed by the increase in bias.

Griffin (1995) investigated these approaches using a two-strata design in the estimation of a proportion of a population with a characteristic of interest. Incidentally, using a two-strata design is quite appropriate to the generalization to multi-strata designs because Kish (1965) observed that the maximum loss in precision occurs when all the remaining weight is subject to the highest weight. Thus, for a fixed ratio of largest to smallest stratum weight, the greatest loss of efficiency occurs in the two strata case. In the present study we will apply this two-strata design to Dual System Estimates (DSE) of census population using empirical data from the 1990 Census Post Enumeration Survey (PES) and the 1995 Test Census ICM.

In general the DSE is equal to the following:

[1] See Wolter (1986); Hogan and Wolter (1988); Mulry and Spencer (1988) for a detailed discussion of dual system estimation and census undercoverage

$$D\hat{S}E = C\frac{(1 - \hat{P}_{ee})}{(1 - \hat{P}_{om})} \qquad (4)$$

where:

C = the census count

$$\hat{P}_{ee} = \frac{1}{\hat{N}_P}[n_1 w_1 \hat{P}_{ee1} + n_2 k w_1 \hat{P}_{ee2}] \qquad (5)$$

$$\hat{P}_{om} = \frac{1}{\hat{N}_E}[n_1 w_1 \hat{P}_{om1} + n_2 k w_1 \hat{P}_{om2}] \qquad (6)$$

where:

$\hat{P}_{ee}$, $\hat{P}_{ee1}$, $\hat{P}_{ee2}$, $\hat{P}_{om}$, $\hat{P}_{om1}$, $\hat{P}_{om2}$

are the estimated proportion of census erroneous enumerations in the total population, in stratum 1 and in stratum 2, respectively, and similarly for census omissions

and,

$\hat{N}_E$, $\hat{N}_p$

are the estimates of the true population count, N, obtained from the E-sample and P-sample, respectively.

Also,

$w_1 = N_1 / n_1$

$w_2 = N_2 / n_2$

$w_2 = kw_1$

where:

k is a constant (the weight ratio)

N_1 and N_2 are the true population counts in strata 1 and 2, respectively , and

$N_1 + N_2 = N$

Consequently,

$$n_1 w_1 + n_2 w_2 = n_1 w_1 + n_2 k w_1 = N \qquad (7)$$

Further, using the first-order Taylor approximation we have the following formulas for the expected value, the variance, and the bias of the dual system estimate:

$$E[D\hat{S}E] \doteq C\frac{(1 - E[\hat{P}_{ee}])}{(1 - E[\hat{P}_{om}])} \qquad (8)$$

where:

$$E[\hat{P}_{ee}] \doteq \frac{1}{N}[n_1 w_1 P_{ee1} + n_2 k w_1 P_{ee2}] \qquad (9)$$

$$E[\hat{P}_{om}] \doteq \frac{1}{N}[n_1 w_1 P_{om1} + n_2 k w_1 P_{om2}] \qquad (10)$$

Next,

$$P_{ee} = \frac{N_1}{N}P_{ee1} + \frac{N_2}{N}P_{ee2} \qquad (11)$$

$$P_{om} = \frac{N_1}{N}P_{om1} + \frac{N_2}{N}P_{om2} \qquad (12)$$

where P_{ee}, P_{ee1}, and P_{ee2} are the true proportion of census erroneous enumerations in the total population, in stratum 1 and in stratum 2, respectively, and P_{om}, P_{om1}, and P_{om2} are defined similarly for census omissions.

$$Var(D\hat{S}E) \doteq C^2[\frac{(1 - P_{ee})}{(1 - P_{om})}]^2[\frac{Var(\hat{P}_{ee})}{(1 - P_{ee})^2} + \frac{Var(\hat{P}_{om})}{(1 - P_{om})^2} - \frac{2Cov(\hat{P}_{ee},\hat{P}_{om})}{(1 - P_{om})(1 - P_{ee})}] \qquad (13)$$

where:

$$Var(\hat{P}_{ee}) = \frac{DEFFw_1^2}{N^2}[n_1 P_{ee1} Q_{ee1} + n_2 k^2 P_{ee2} Q_{ee2}] \quad (14)$$

$$Var(\hat{P}_{om}) = \frac{DEFFw_1^2}{N^2}[n_1 P_{om1} Q_{om1} + n_2 k^2 P_{om2} Q_{om2}] \quad (15)$$

$$Cov(\hat{P}_{ee}, \hat{P}_{om}) = \frac{-DEFFw_1^2}{N^2}[n_1 P_{ee1} P_{om1} + n_2 k^2 P_{ee2} P_{om2}] \quad (16)$$

where DEFF is the design effect due to cluster sampling for the estimated erroroneous enumeration and omission rates. In the present study, the DEFF is a simple average of the design effects for the erroneous enumeration rate and the omission rate in stratum 1. Therefore DEFF = ($DEFF_{ee1}$ + $DEFF_{om1}$) / 2. In addition, $Q_{eeh} = 1 - P_{eeh}$, and $Q_{omh} = 1 - P_{omh}$.

Also,

$$Bias(D\hat{S}E) = E[D\hat{S}E] - Target \quad (17)$$

where,

$$Target = N = C\frac{(1 - P_{ee})}{(1 - P_{om})} \quad (18)$$

Finally,

$$MSE(D\hat{S}E) = Var(D\hat{S}E) + Bias(D\hat{S}E)^2 \quad (19)$$

In the simulation conducted for the present study, each of the values used were taken from actual parameter values encountered in the 1990 PES or 1995 DSE. Following Cox and McGrath (1981) and Griffin (1995), an estimated mean square trimming analysis (MSE) was conducted using a two stratum design. First, a pair of P_{ee} and P_{om} were alternatively set to actual values from each of the eight strata in Sacramento. For each pair of P_{ee} and P_{om} values set from a particular stratum, two pairs of simulated values, P_{ee1}, P_{om1}, P_{ee2}, P_{om2}, were derived from the P_{ee} and P_{om}. In addition, for each of the eight Sacramento strata, a simulated target population value, N, was set using formula (18). Values for N_1 and N_2 were then both set as 0.5*N. Similarly the value for n was taken from the particular stratum value. The simulated values for n_1 and n_2 were then calculated using a constant parameter value for the unbiased weight ratio, k, and the values N and N_1 calculated above and inserted into formula (7). The values for w_1 and w_2 were then set. The DEFF was set as explained in the discussion that follows formula (16).

Finally, simulated values for :

$$Var(\hat{P}_{ee}),\ Var(\hat{P}_{om}),\ Cov(\hat{P}_{ee},\ \hat{P}_{om}),$$
$$E[\hat{P}_{ee}],\ E[\hat{P}_{om}],\ E[D\hat{S}E],$$
$$Var(D\hat{S}E),\ Target,\ Bias(D\hat{S}E),\ MSE(D\hat{S}E)$$

were then calculated using the values set above and inserted into formulas (14), (15), (16), (9), (10), (8), (13), (18), (17), and (19), respectively.

In all for the present study, for each of the eight simulated pairs of 'true' values for P_{ee} and P_{om} as well as for N, N_1, N_2, n, n_1, n_2, and for DEFF, eight additional sets of simulated 'true' values for P_{ee1}, P_{om1}, P_{ee2}, P_{om2} (one for each of the eight strata) were calculated resulting in 64 combinations of simulated values inserted into the variance, covariance, expected value, bias and MSE estimators listed above.

Next, for each of the following :

$$Var(D\hat{S}E),\ Bias(D\hat{S}E),\ MSE(D\hat{S}E)$$

a simple average of the 64 values was calculated for the each of the series of unbiased weight ratios, k = 25, 20, 25, 10, and 5. For each of these unbiased weight ratios, k, the biased weight ratio, j, was successively decremented by 1 and replaced in k (creating a new set of 64 variance and bias values) until the minimum average MSE was obtained. The results showing the optimal trimmed weight ratios resulting for each unbiased weight ratio, k, are shown in Table 1.

For each of the unbiased weight ratios, the plot of average MSE values by trimmed (biased) weight ratio formed a "U" pattern. As the weight ratio was trimmed from the unbiased weight ratio, the variance decreased and the bias increased so that the average MSE decreased to the optimal point and then began rising again. This result was consistent with the pattern found by Griffin, 1995.

Calculation of Shrinkage Weights

A second method employed for determining the most favorable weight was the shrinkage method (Stokes, 1989; Griffin, 1995).

Stokes(1989) puts forth that the optimal weights are a weighted average of the unbiased sampling weight in stratum i, $w_i = N_i/n_i$, and the proportional sampling weight, N/n. The optimal weights are determined by the relative variability of the estimator within each stratum with that of the estimator between the strata. This value is characterized by B which is calculated as follows :

$$B = \frac{\frac{\tau_i^2}{n_i}}{[\sigma^2 + \frac{\tau_i^2}{n_i}]} \qquad (20)$$

where,

τ_i^2 = population variance within stratum i

σ^2 = population variance between strata

n_i = sample size in stratum i

In the present study,

In the present study the following formulas were used to calculate the stratum i shrinkage weight, w_{si}, for each unbiased weight, w_i = 25, 20, 15, 10, and 5:

$$w_{si} = B_i(\frac{N}{n}) + (1 - B_i)w_i \qquad (21)$$

$$B_i = \frac{WithVar_i}{WithVar_i + BetVar} \qquad (22)$$

$$WithVar_i = \frac{1}{n_i}(P_{eei}Q_{eei} + P_{omi}Q_{omi}) \qquad (23)$$

$$BetVar = \frac{1}{4}((P_{ee1} - P_{ee2})^2 + (P_{om1} + P_{om2})^2) \qquad (24)$$

The ratio the stratum shrinkage weights, w_{s2} / w_{s1} for each unbiased weight, was then calculated and averaged for the same 64 combinations of values of P_{ee1}, P_{ee2}, P_{om1}, and P_{om2} as in the MSE analysis (Table 1.)

Table 1. Unbiased and Optimal Trimmed Weight Ratios Using Minimum MSE and Shrinkage Approaches

Unbiased Weight Ratio	Biased Weight Ratio with Minimum Average MSE	Shrinkage Weight Ratio
25	10	17.82
20	9	14.70
15	8	11.42
10	6	7.94
5	4	4.22

Discussion

The discrepancy found between the results of the minimum MSE and shrinkage approaches may be due to the large emphasis put on the unbiased weight in the shrinkage weight formula. Our preference would be the more direct approach of the minimum MSE technique.

This is also more consistent with the anticipated approach to the research that will be done on weight variation following the 1998 Dress Rehearsal ICM.

Conclusion

If weight trimming is called for in the Census 2000 ICM program, this research along with that of Griffin (1995) provides efficient values to which to trim the most extreme weights. Further research will concentrate on an empirical investigation of the 1998 Dress Rehearsal ICM results to see if which of our findings in the present study are corroborated. Weight trimming is not expected to be implemented in the Dress Rehearsal ICM but is a possibility for the 2000 Census ICM.

References

Cochran, W.G. (1977), *Sampling Techniques*, Wiley.

Cox, B.G. and McGrath, D.S. (1981), "An examination of the effect of sample weight truncation on the mean square error of survey estimates", Presented at the Biometrics Society ENAR meeting, Richmond, VA. March 1981.

Hogan, H. and Wolter, K. (1988), "Measuring Accuracy in a Post-Enumeration Survey", *Survey Methodology*, 14, 99-116.

Griffin, R. (1995), "Dealing with Wide Weight Variation in Polls", Proceedings of the Survey Research Methods Section of the American Statistical Association, 908-911.

Kalton, G. (1983), *Introduction to Survey Sampling*, Sage.

Kish, L. (1965), *Survey Sampling*, Wiley.

Mulry, M.H. and Spencer, B.D. (1988), "Total Error in the Dual System Estimator: The 1986 census of central Los Angeles County", *Survey Methodology*, 14, 241-263.

Potter, F. (1988), "Survey of Procedures to Control Extreme Sampling Weights", Proceedings of the Survey Research Methods Section of the American Statistical Association, 453-458.

Schindler, E. (1998), "Weight Variation in the ICM Program", Internal Census Bureau memorandum.

Stokes, L. (1989), "Improvement of Precision by Shrinking of Sample Weights", Draft Paper.

Wolter, K. (1986), "Some Coverage Error Models for Census Data", *Journal of the American Statistical Association*, 81, 338-346.