SAMPLE ALLOCATION RESEARCH FOR THE CENSUS 2000 ICM SURVEY

Richard A. Griffin and Felipe Kohn
U.S. Bureau of the Census

Key Words: poststratification, allocation, multipurpose optimization

INTRODUCTION

The Census 2000 Integrated Coverage Measurement (ICM) Survey will be used to provide estimated census totals that correct for the undercount, especially the differential undercount among racial, ethnic, and socioeconomic groups, that has been observed in every decennial census from 1940 onward. The ICM survey will be designed to produce direct estimates of total population for each of the 50 states, the District of Columbia (DC), and Puerto Rico and will have a sample size of about 750,000 housing units (HUs) excluding DC and Puerto Rico. This paper will present results of research on methods to allocate the ICM sample within a state.

Design issues discussed include sampling stratification, poststratification, sample allocation, and the resulting precision of Dual System Estimates (DSE) of the total population as well as for demographic subgroups. Optimal allocations for total population and for major poststrata groups, as well as proportional allocation, will be studied. Additional allocation schemes considered are: (1) two methods of allocation with more than one item of interest suggested in Cochran (1977); (2) the multipurpose optimization suggested by Kish (1987); and (3) another optimal allocation scheme for multiple response variables suggested by Rahim and Currie (1993). These last two methods involve assigning relative weights of importance to major estimates and minimizing loss functions for a given set of weights using Lagrange multipliers. A number of sets of weights are considered. Results from the 1995 Census Test in Oakland, California and Paterson, New Jersey are used to simulate and evaluate these alternative allocation schemes.

To measure the coverage and correct for coverage errors two samples are needed. The same blocks are included in the P (population) sample and the E (enumeration) sample. The E sample consists of all initial enumerations, correct or incorrect, in the sample blocks. The P sample consists of all people determined by the ICM interview to have been living in the housing units at the time of the initial Enumeration. The joint implementation of these two samples constitutes the ICM survey.

Note: 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.

For the 1995 Census Test in Oakland and Paterson a stratified sample of block clusters was selected with the objective of producing reliable estimates of the population for various groups defined by race/tenure cross-classified by sex and age. The post nonresponse follow up counts in the sample block clusters constitute the E sample and the ICM enumeration results in these same sample block clusters make up the P sample. In Oakland, 150 block clusters averaging about 69 HUs were selected resulting in a sample of about 24,000 persons. In Paterson, 100 block clusters averaging about 65 HUs were selected resulting in a sample of about 20,000 persons.

The ICM sample is designed to provide sufficient precision for the dual system estimates of total population for the ICM poststrata. The term "poststrata" is used to denote the finest level of detail for which ICM estimates will be produced. The poststrata are defined by characteristics of the persons enumerated in the ICM and are as homogeneous as possible with respect to the census undercount mechanisms. For the 1990 Post Enumeration Survey (PES) the variables used to define poststrata were Census division, size and type of place, race, origin and overall population of the poststrata. Subsequently, the poststrata were further partitioned by age, sex, and in some cases tenure (owner, renter). Details of the 1990 PES sample design are given in Woltman et al. (1988).

The major poststrata for the Oakland ICM were as follows:
1. Non-Hispanic White and Other renters
2. Non-Hispanic White and Other owners
3. Black renters
4. Black owners
5. Non-Black, Non-Asian and Pacific Islander (API) Hispanic renters
6. Non-Black, Non-API Hispanic owners
7. API renters
8. API owners

The major poststrata for the Paterson ICM were as follows:
1. Non-Hispanic White and Other renters
2. Non-Hispanic White and Other owners
3. Black renters
4. Black owners
5. Non-Black Hispanic renters
6. Non-Black Hispanic owners

As described in Navarro (1995) the goal of the 1995 Census Test ICM sample design was to develop sampling strata to support estimation for major poststrata defined by race/origin and tenure. This was accomplished by creating sampling strata with high concentrations of the race/origin/tenure groups corresponding to the major poststrata. For both Oakland and Paterson, several algorithms for defining sampling strata were analyzed examining proportional allocation and optimal allocation for the major race/origin groups. For each allocation, the resulting reliability levels on the estimated number of persons missed in the census for the major race/origin groups were calculated. For example, for the selected Oakland stratification scheme, if the sample was optimally allocated for Blacks, then the variance for the estimated number of missed Blacks would have been about 89 percent of what would of been the result under proportional allocation. Remember, that while for Black estimates a reduction in variance would have been realized, the accuracy of estimates for other race/origin groups and the total population may have deteriorated. For both Oakland and Paterson, a decision was made to allocate the total sample proportional to the size of each sampling stratum, that is, proportional allocation. The following sampling stratification schemes were used:

Oakland:
1. Black - Block clusters with 40 percent or more Blacks
2. Hispanic - Block clusters with less then 40 percent Black and 10 percent or more Hispanic
3. Asian and Pacific Islander (API) - Block clusters with less than 40 percent Black, less than 10 percent Hispanic, and 15 percent or more API
4. Other - All remaining block clusters

Paterson:
1. Black - Block clusters with 30 percent or more Blacks
2. Hispanic - Block clusters with less than 30 percent Black and 10 percent or more Hispanic
3. Other - All remaining block clusters.

Dual system estimates were calculated by race/origin/tenure/age/sex groups. Components of these estimates (such as weighted erroneous enumerations from the E sample and weighted matches from the P sample) were summed over the age/sex groups for input to the analysis for this paper.

METHODOLOGY

For a given poststrata k, the dual system estimate (DSE) of population is given by:

$$DSE_k = C_k(1-P_{ek})/(1-P_{ok})$$

where
C_k = the census count in poststratum k.
P_{ek} = the weighted proportion of erroneous enumerations in poststratum k.
P_{ok} = the weighted proportion of omissions in poststratum k.

The variance of DSE_k was calculated using Taylor Linearization, assuming erroneous enumeration rates, omission rates, and design effects do not vary by sampling stratum within poststratum. It was also assumed a person can't be both an omission and an erroneous enumeration. Sampling fractions were ignored since they are not a factor for sample allocation to sampling strata.

In this study simple random sample variances for estimated proportions were used with design effects to account for the fact that the ICM sample is a sample of block clusters not a sample of persons. For major poststrata defined by race/origin and tenure from the 1995 Census Test in Oakland and Paterson, variances of estimated erroneous enumeration rates (from the E sample) and omission rates (from the P sample) were calculated by a Jackknife procedure dropping out one block cluster at a time with no reweighting. These variances were divided by simple random sampling PQ variances to produce a design effect for erroneous enumerations and a design effect for omissions for each major poststratum.

The variance of the DSE for a poststratum k is given by

$$VAR(DSE_k) = N_k^2 S_k^2 \sum_h T_{hk}^2 / n_h A_{hk}$$

Where h denotes the sampling stratum. N_k is the population in poststratum k, S_k^2 is a term for poststratum k that accounts for the variance of the estimated omission rate and erroneous enumeration rate as well as the covariance between these rates and uses design effects, n_h is the person sample size allocated to poststratum k , and A_{hk} is the proportion of persons in sampling stratum h who are in major poststratum k.

Now summing over the major poststrata the overall DSE is given by

$$DSE = \sum_k DSE_k$$

with

$$VAR(DSE)=\sum_h\sum_k N_k^2 S_k^2 T_{hk}^2/n_h A_{hk}$$

ignoring the covariance terms.
Define

$$B_h=\sum_k N_k^2 S_k^2 T_{hk}^2/A_{hk}$$

Using the above notation, we can describe the following allocations to the sampling strata for each site (Note: Optimal allocations throughout this paper are determined by minimizing a function, such as the variance of an estimate, using Lagrange multipliers).
n is the total sample for each site; n = 24,227 for Oakland and n = 19,617 for Paterson
1. Proportional allocation:

$$n_h=nW_h$$

where W_h is the proportion of the 1995 test Census persons in a site (Oakland or Paterson) in sampling stratum h.
2. Optimal for major poststratum group k:

$$n_h=n\frac{T_{hk}/\sqrt{A_{hk}}}{\sum_h T_{hk}/\sqrt{A_{hk}}}$$

3. Optimal for overall DSE

$$n_h=n\frac{\sqrt{B_h}}{\sum_h\sqrt{B_h}}$$

4. Average of the optimal allocations over the estimates for all the major poststrata and the overall DSE. This is suggested in Cochran (1977) as a satisfactory compromise for the problem of allocation with more than one estimate considered important. Denote each of these estimates by the subscript k and let nhko be the optimal allocation to sampling stratum h for estimate k. Denote the number of estimates by K. Then:

$$n_h=\frac{\sum_k^K n_{hk}^o}{K}$$

5. An alternative compromise suggested by Chatterjee (1967) and cited in Cochran (1977) is to choose the nh to minimize the average of the relative increases in variance resulting from deviations from the optimal taken over all the estimates of interest. This allocation is as follows using the notation from allocation 4 above:

$$n_h=n\frac{\sqrt{\sum_k n_{hk}^{o2}}}{\sum_h\sqrt{\sum_k n_{hk}^{o2}}}$$

6. Kish (1987) discusses multipurpose designs and suggests averaging between all the optimal allocations by minimizing the combined weighted variance for a fixed sample size. This approach involves assigning relative values of importance to all the estimates of interest. Let V2k(min) denote the minimum variance attainable with optimal allocation of the sample for the estimate k. In general write

$$VAR(DSE_k)=\sum_h V_{hk}^2/n_h$$

Thus

$$1+L_k(n_h)=\frac{\sum_h V_{hk}^2/n_h}{V_{k(\min)}^2}$$

is the ratio of increase in variance with the allocation n_h to the minimum variance for estimate k. $L_k(n_h)$ is the relative loss over the minimal value 1. Kish proposes an average loss function for any set of allocations n_h of the sample, where the loss for each estimate k is weighted with a factor I_k assigned for its relative importance.

$$1+L(n_h)=\sum_k I_k(1+L_k(n_h))$$

Which equals

$$\sum_h Q_h^2/n_h, \text{ where } Q_h^2=\sum_k I_k \frac{V_{hk}^2}{V_{k(\min)}^2}$$

This can be minimized by the allocation

$$n_h=n\frac{Q_h}{\sum_h Q_h}$$

7. Rahim and Currie (1993) suggest dealing with multiple response variables by minimizing a distance function defined as a weighted sum over the estimates of interest of their relative variances (CV^2).

$$D = \sum_k I_k CV^2(DSE_k) = \sum_h \sum_k \frac{I_k V_{hk}^2}{DSE_k^2 n_h}$$

Which equals

$$\sum_h F_h^2/n_h \text{ where } F_h^2 = \sum_k \frac{I_k V_{hk}^2}{DSE_k^2}$$

D is minimized by the following allocation:

$$n_h=n\frac{F_h}{\sum_h F_h}$$

RESULTS

Our summary statistic for a given allocation is the sum over the major poststratum estimates of the relative difference between the CV for that allocation and the optimal allocation CV for the estimate.

$$SumRelDiff=\sum_k \frac{(CV_{jk}-CV(OPT)_k)}{CV(OPT)_k}$$

SEE TABLE 1 AT THE END OF THE PAPER

For Oakland the sum of relative differences from optimal is 0.541 using the Chatterjee allocation, 0.621 using the optimal allocation for the total DSE estimate, 1.609 using the optimal allocation for the black owner estimate, the highest sum for Oakland, and 0.591 using proportional allocation as was done in the 1995 test.

SEE TABLE 2 AT THE END OF THE PAPER

For Paterson, the sum of relative differences from optimal is 0.509 using the Chatterjee allocation, 0.779 using the optimal allocation for the total DSE, 1.471 using the optimal allocation for the black renter estimate, the highest for Paterson, and 0.512 using proportional allocation.

The Chatterjee allocation minimizes the proportional increase in variance (assuming sampling fractions are negligible) resulting from deviations from the optimal allocation. Thus, as we would expect, for Oakland and Paterson the Chatterjee allocation is the best (lowest) using this sum as a measure. Note, however, that this sum might not be the best measure for selection of an allocation since it implicitly gives equal weight to each estimate.

The following observations can be made from Tables 1 and 2.

- Optimal allocations for individual major poststratum groups are significantly worse using this summary statistic.
- The Chatterjee, Average Optimal, and Proportional allocations are the best and give similar results.
- The Kish and Rahim and Currie multipurpose optimal allocations do well for 2 sets of weights; 1) equal weights to all estimates, and 2) 0.5 weight for the overall DSE and the rest of the weight evenly distributed.

CONCLUSIONS AND RECOMMENDATIONS FOR CENSUS 2000

For the 1995 Census Test proportional allocation to the sampling strata was used. For the 1990 PES, optimal allocation for the total population estimate was used. For this research, using the assumptions of this paper, the Chatterjee, average optimum and proportional allocation methods are the best allocations among those that do not require assigning importance weights to the estimates of interest. "Best"is measured by the sum of the relative differences in CV from optimal over the estimates. This measure implicitly assigns equal weights to the estimates. All three of these allocations provide very close to the optimal CV on the total population estimate. The Kish and Rahim and Currie allocation methods depend on the assignment of importance weights. Assigning half the weight to the total population estimate and the rest of the weight evenly to the other major poststratum estimates, as well as assigning equal weights to all the estimates do best by our quality measure. Since all the major poststrata, as

Well as the total population estimate, are important for the use of ICM in Census 2000, the Chatterjee allocation appears to be the best choice. However, since the proportional allocation method is much easier to explain and gave results very close to the Chatterjee allocation, the proportional allocation may make more sense. The differences between the two allocations are probably less than the error caused by the assumptions used in achieving the results.

For Census 2000, we will need to form sampling strata and allocate the sample to these strata for each of the 50 states as well as the District of Columbia and Puerto Rico. Using the best available data from the 1990 PES, the 1995 Census Test and the Census 2000 Dress Rehearsal, we plan on examining all the allocation methods in this paper to determine what is best for the Census 2000 ICM sample design. We will also look at numerous ways to define the sampling strata. For example, for the Census 2000 Dress Rehearsal we added tenure to the race/origin sampling strata used in the 1995 test (we know tenure is highly correlated with undercount) and looked at various cutoff points (i.e., Blocks with more than 30% Black renters or Blocks with more than 40% Black renters).
For Census 2000 research we will also look at "hard-to-count" scores which have been developed by tract from the 1990 Census as a possible factor in defining sampling strata. In addition, research is ongoing at the Census Bureau on refinement of poststratification for estimation to reduce heterogeneity bias. Hard-to-count scores, inclusion probabilities estimated using logistic regression, and raking are the leading ways being considered to improve poststratification. For ICM sample design research on forming sampling strata and allocating the sample to sampling strata for Census 2000, we will want to look at the effect on poststrata defined as indicated from the results of this research. We will also look at the effect of the ICM sample design on small area estimates.

REFERENCES

Chatterjee, S. (1967), "A Note on Optimum Stratification", Skand. Akt, Volume 50, 530-534.

Cochran, W.G. (1977), Sampling Techniques, John Wiley & Sons Inc.

Kish, L. (1987), Statistical Design for Research, John Wiley & Sons Inc.

Navarro, A. and Woltman, H. (1995), "1995 Census test: Integrated Coverage Measurement Sample Design," Proceedings of the Section on Survey Research Methods, American Statistical Association, 718-723.

Rahim, M.A. and Currie S. (1993), "Optimizing Sample Allocation for Multiple Response Variables," Proceedings of the Section on Survey Research Methods, American Statistical Association, 346-351.

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

TABLE 1 - OAKLAND			
		SUM OF	
		RELDIFF	
ALLOCATIONS		OF CV FROM	
OPTIMAL FOR		OPT CV	
NON-HISPANIC WHITE	RENTER	0.941	
AND OTHER	OWNER	1.407	
BLACK	RENTER	1.354	
	OWNER	1.609	
NON-BLACK, NON-API HISPANIC	RENTER	1.205	
	OWNER	0.881	
API	RENTER	0.939	
	OWNER	0.770	
TOTAL DSE		0.621	
AVERAGE OPT		0.546	
CHATTERJEE		0.541	
PROPORTIONAL		0.591	
WEIGHT SETS		KISH	RAHIM
1		0.551	0.612
2		0.541	0.614
3		0.843	0.783
4		0.746	0.824
5		0.640	0.692

Table 2 - Paterson			
		Sum of	
		RELDIFF	
ALLOCATIONS		of CV from	
Optimal for		opt CV	
Non-Hispanic White	Renter	1.175	
and Other	Owner	1.241	
Black	Renter	1.471	
	Owner	1.396	
Non-Black Hispanic	Renter	0.992	
	Owner	0.747	
Total DSE		0.779	
Average OPT		0.517	
Chatterjee		0.509	
Proportional		0.512	
Weight Sets		KISH	RAHIM
1		0.543	0.711
2		0.510	0.660
3		1.031	1.062
4		0.654	0.672
5		0.540	0.556