1995 CENSUS TEST: INTEGRATED COVERAGE MEASUREMENT SAMPLE DESIGN

Alfredo Navarro, and Henry F. Woltman
U.S. Bureau of the Census
Alfredo Navarro, Bureau of the Census, Washington, D.C. 20233

I. Introduction

A major objective of the 1995 Test Census is to develop a new methodology for coverage estimation, referred to as Integrated Coverage Measurement or simply ICM. The main objective of the ICM system is to reduce differential undercount. The basic assumption underlying the ICM design is the existence of a single best census number obtained as a result of incorporating counting and estimation methods. Counting refers to techniques by which direct contact with respondents is attempted, such as mail, telephone, personal visit, or by other means. Estimation refers to the use of statistical techniques to develop estimates for persons or units not contacted by the more traditional counting techniques. The 1995 Test Census is also testing a variation of the nonresponse followup operation by which only a sample of the nonresponse population is contacted by personal visit. The combination of counting and estimation resulting from conducting nonresponse follow-up on a sample basis and the ICM survey will be the basis for the "one number" census estimate for the 1995 Test Census.

Two estimators of population size will be calculated in the 1995 Test Census. The first estimator is the usual dual system estimator or DSE analogous to capture-recapture estimation for wildlife populations. The second estimator relies on a second collection effort, conducted in a probability sample of blocks, to obtain the best possible count of actual persons in the sample areas. Reconciliation between the census and the second enumeration results in the final estimate based on ratio estimation which we will call censusplus estimation.

This paper provides an overview of the design, size of the sample and expected standard errors of population size due to sampling the nonrespondents and for coverage measurement. Design issues that are discussed include stratification, sample allocation and expected measures of reliability of census plus estimates for various demographic subgroups of the population. Stratification and poststratification are discussed in Section II. Section III.A describes estimators of population size. Section III.B gives an expression of the variance. Section III.C describes an analytical method used to determine the ICM sample size and the statistical methodology used in a second simulation designed to estimate the unconditional or "total" sampling error of population estimates, including that introduced by sampling the non-mail return population. Section IV discusses results and sample size recommendations.

II. Sampling Strata and Poststrata.

A. Poststrata

The objective of the coverage measurement survey component of the 1995 Census Test is to produce estimates of the population for various groups defined by race/tenure cross-classified by sex and age. These groups are referred to as poststrata. The poststrata define the population groups for which direct estimates of population size will be produced. A description of the poststrata is given below.

Race/Origin (4)
Non-Hispanic White and Other
Black (African American)
Non-Black Non-API Hispanic
Asian and Pacific Islander

Tenure (2)
Owner
Non-owner

Age/Sex (7)

0 to 17	Male and Female
18 to 29	Male
18 to 29	Female
30 to 49	Male
30 to 49	Female
50 - over	Male
50 - over	Female

For Oakland we will be able to provide reliable estimates for all poststrata. The Asian and Pacific Islander population in Paterson is not large enough to support API poststratification. The Hispanic and API populations in NW Louisiana are very small, therefore reliable estimation for groups other than Black and Non-Black is not possible. Thus, direct estimation of the Hispanic and API was not recommended.

B. Sampling Strata

The goal of the ICM sample design is to develop sampling strata to support estimation as defined above.

This is accomplished by creating sampling strata with high concentrations of the race/tenure groups corresponding to the poststrata.

1. Race Stratification by Block

The geographic units used for stratification were 1990 census block. The use of 1990 information (for stratification), although not perfect, should result in an adequate and effective stratification to improve the estimates, particularly for minorities. Note that the main motivation for racial stratification in the 1995 Census Test is to insure that target groups are adequately represented in the ICM sample. For example, for Oakland, blocks having more than 40 percent Black were placed in the same sampling stratum.

To improve the stratification for Asian and Pacific Islanders (API) in Oakland we tried several algorithms to assign blocks to strata. The first algorithm used created the Black sampling stratum first, followed by the Hispanic and API. All remaining blocks were grouped together to form the Non-Hispanic White and Other sampling stratum. The implementation of this algorithm resulted in a 29 percent "enrichment factor"[1] for the API sampling stratum. Thirty-five (35) percent of the API population were contained in the API sampling stratum. To improve these factors we tried a second algorithm which created the API sampling stratum first, followed by the Black and Hispanic. The results for the API were significantly better, an "enrichment factor" of 38 percent and 63 percent of the API population were contained in the sampling stratum. Further analysis revealed the improvement produced a significant reduction in the size of the Hispanic sampling stratum. The size of the Hispanic sampling stratum was reduced to less than half of the original size while containing less than one-third (originally, almost 50 percent) of the Hispanic population. Note that Hispanic and Asian Americans tend to live together in higher proportions than with Blacks. The effect of Hispanic and Asian stratification on variances may be marginal. After analyzing the results of several algorithms, we decided to base our recommendation on results from the first algorithm.

2. Comparison Criteria

To evaluate and compare the effectiveness of several stratification schemes, the following statistical model was used.

For each group i, define an estimate of the proportion P of persons with characteristic X (e.g. missed by the census), as follows:

$$\hat{P}_i = \sum_h \frac{N_{hi}}{N_i} \hat{P}_{hi} \quad \text{with} \quad \hat{P}_{hi} = \frac{\hat{X}_{hi}}{\hat{N}_{hi}}.$$

$\hat{P}_{hi}$ is the proportion of persons with characteristic X in group i and stratum h. $\hat{P}_i$ is a separate ratio estimate.

Let

$\alpha_i = N_i/N$ be the proportion in group i, say proportion of Black,

$\alpha_{hi} = N_{hi}/N_h$ be group i proportion in stratum h, and

$W_h = N_h/N$ be the relative stratum size. Then

$$Var(\hat{P}i) \doteq \sum_h \frac{\alpha_{hi}^2}{\alpha_i^2} W_h^2 \frac{P_{hi}Q_{hi}}{n_h \alpha_{hi}} \quad (1)$$

Recall $E[n_{hi}] = n_h * N_{hi}/N_h$.

If $P_{hi} = P_i$ for all h,

then $$Var(\hat{P}_i) = \frac{P_i Q_i}{\alpha_i^2} \sum_h \alpha_{hi} \frac{W_h^2}{n_h} \quad (2)$$

- For **proportional allocation**, that is,

$n_h = n * W_h$

$$\Rightarrow \quad Var(\hat{P}_i) = \frac{P_i Q_i}{\alpha_i^2 n} \sum_h \alpha_{hi} W_h . \quad (3)$$

- For **optimum allocation**, that is,

$$n_h = n\left(\frac{W_h\sqrt{\alpha_{hi}}}{\sum_h W_h\sqrt{\alpha_{hi}}}\right) \Rightarrow$$

$$Var(\hat{P}_i) = \frac{P_i Q_i}{\alpha_i^2 n} \left[\sum_h W_h\sqrt{\alpha_{hi}}\right]^2 \quad (4)$$

The ratio R(Opt/Prop) = $$\frac{\left[\sum_h W_h\sqrt{\alpha_{hi}}\right]^2}{\sum_h \alpha_{hi} W_h}$$

reflects the reduction in variance of "optimum" allocation over proportional allocation ($R \leq 1$.)

For instance, if the following stratification is used for Oakland

Black - 30 percent of more Black
population,

Hispanic - less than 30 percent Black and 10 percent
of more Hispanic,

API - less than 30 percent Black, less than 10 percent Hispanic, and 10 percent or more API,
Other - remainder

If the sample is "optimally" allocated for Black, then the variance is about 89 percent of what would result under proportional allocation. Remember, that while for Black estimates a reduction in variance is realized, the accuracy of estimates of characteristics of the total population may deteriorate. The following stratification scheme was used for ICM for the 1995 Test.

OAKLAND

Black - more than 40 percent
Hispanic - 40 percent or less Black and 10 percent or more Hispanic
API - less than 40 percent Black, less than 10 percent Hispanic, and 15 percent or more API
Other - remainder

PATERSON

Black - more than 30 percent
Hispanic - 30 percent or less Black and 10 percent or more Hispanic
Other - remainder

NW LOUISIANA

Black - more than 20 percent
Other - remainder

The sample was allocated proportional to the size of the stratum based on the 1990 census population counts.

III. Variance of the Population Size Estimate

A. Estimators of Population Size

Two estimators of population size will be calculated in the 1995 Census Test sites. The estimators are the dual system estimator (DSE) or CensusPlus (C+) estimator.

1. Dual System Estimate

Define for a particular poststratum j:

C_j - census total population count
I_j - count of persons whose charac-teristics are entirely missed in the census

$\hat{EE}_j$ - estimated number of erroneous enumerations from ICM
$\hat{N}_j$ - estimated population size from ICM
$\hat{M}_j$ - ICM estimate of matches. A "match" is a person found in both enumerations, the census and the second enumeration.

$$D\hat{S}E_j = (C_j - I_j - \hat{EE}_j) \frac{\hat{N}_j}{\hat{M}_j}$$

The DSE estimate of total population is

$$\hat{P} = \sum_j D\hat{S}E_j$$

2. Censusplus Estimate

Let $\hat{C}+$ be the estimate of total population.

with

$\hat{\beta}_j$ - estimated adjustment factor for poststratum j

$$\hat{\beta}_j = \frac{\hat{C}_j^+}{\hat{C}_j} \quad \text{where}$$

$\hat{C}_j^+$ - estimated census plus count for poststratum j.

This is the weighted sum of the "resolved" population counts across all sample blocks. The "resolved" block counts are the result of the second enumeration effort and the initial census enumeration.

$\hat{C}_j$ - estimated census count from ICM sample.

The C+ estimate of total population is

$$\hat{C}^+ = \sum_j \beta_j \sum_{b\epsilon Site} C_{bj}$$

The second sum in the above formula is an estimate if the nonrespondents are sampled.

B. Variance of Estimator

The variance of the population estimate as described above is as follows.

$$Var(\hat{C}+) = Var(\sum_j C_j \beta_j) =$$

$$\sum_j C_j^2 Var(\beta_j) + \sum_j \sum_{i \neq j} C_i C_j Cov(\beta_i, \beta_j)$$

Assuming a simple random sample of blocks (block clusters)[2] without replacement the variance of the adjustment factor is as follows.

$$Var(\beta_j) \doteq (1 - \frac{b}{B}) \frac{S_j^2}{b\bar{C}_j^2}$$

where

B - total blocks in site
b - number of blocks in the ICM sample
C_j - average census population per block in poststratum j

$$\bar{C}_j = \frac{\sum_{b=1}^{B} C_{jb}}{B} \quad \text{and} \quad S_j^2 = \frac{\sum_{b=1}^{B} (C_{jb}^+ - \beta_i C_{jb})^2}{B-1}$$

The covariance between the j-th and i-th adjustment factors is as follows.

$$Cov(\hat{\beta}_j, \hat{\beta}_i) \doteq (1-\frac{b}{B})\frac{S_{ij}}{b\bar{C}_i\bar{C}_j}$$

C. Statistical Simulations

The CensusPlus estimate of total population, as simulated from the 1990 PES data base is as follows.

$$\hat{C}+ = (\hat{C}-E\hat{E}) + (\hat{N} - \hat{M})$$

This is the numerator of the adjustment factor as defined in III.A.2. The second term in the right hand side of the equation is an estimate of the number of persons missed by the census. This is precisely the component of the population that potentially will be found by the second enumeration effort or censusplus. Schindler and Navarro [2] simulated the variance of the C+ estimate using the 1990 PES data base and found that for a fixed sample size the variance of the DSE and C+ estimates are about the same. Since the variances are similar we decided to simulate the expected variance of the C+ estimates for the sites. We used the 1990 PES database to simulate the C+ estimates and its variance.

1. Analytical Simulation.

This section describes the statistical methodology used to simulate the C+ estimate and its expected variance for various population subgroups. This simulation does not include the additional sampling error due to NRFU sampling. The objective of the simulation is to approximate the expected variance of $\hat{C}$+.

This problem can be approached in two different ways.

a.) Tamper with the PES data so that it looks like the site data. Use the transformed data set to develop measures of uncertainty and sample size, or

b.) Simulate the C+ estimate based on the 1990 PES results and approximate the variance of $\hat{C}$+ . Adjust the variance of $\hat{C}$+ to account for differences in demographics between the PES and the sites.

We favored the second approach. Causey [6] performed a simulation as described in a.) and obtained measures of uncertainty and sample sizes very similar to our results.

To perform the simulation we obtained values for $\hat{\beta}_j$, S_j^2 , and S_{ij} (or the correlation matrix) for each poststratum (as defined in Section II.A). For the simulation we assumed a simple random sampling design for ICM.

Up to eight population groups or poststrata were used per site. Basic groups were defined by Black owners and renter, Asian and Pacific Islander owners and renters, non-black non-API Hispanic owners and renters, and all other owners and renters. Because of their small numbers, American Indians and APIs (except in Oakland), were included with the Non-Hispanic White/Other populations. In NW Louisiana only two groups were defined, Black and Non-Blacks. Oakland was considered a large urban area, Paterson a small urban area, and the Louisianan site a nonurban area. For each of the sites we used the PES data from the appropriate poststratum groups. All Oakland poststrata except API and the Other poststrata for the other sites are defined for the region. For APIs in Oakland and all minority populations in Paterson and Louisiana the poststrata are national. Limiting the PES input data to those portions of the poststrata in the same region as the test site would have reduced the standard errors for Paterson by about one third. From a design perspective it was better to assume the larger standard errors.

For example, to simulate the adjustment factor for Black owners in Oakland we used the 1990 PES poststratum number 28 - Black owners in urbanized areas with more than 250,000 population in the West.

Population variances and covariances were approximated as follows.

$$\hat{S}_j^2 = \frac{\sum_{m=1}^{M_j}(C_{jm}^{+} - \beta_j C_{jm})^2}{M_j-1} \text{, where}$$

M_j - number of PES blocks with persons in the poststratum.

C_{im} and C_{im}^{+} are the census and censusplus counts for block m, respectively. The censusplus count is sometimes referred to as the "resolved enumeration".

The population variances were adjusted to be used at the site level. The adjustment was necessary to take into account that a proportion of the block clusters in the site have no persons in some of the poststrata. Each of the population variances were adjusted by the ratio B_j/B, B_j is the number of blocks with persons in the j-th poststratum and B is the total number of blocks in the site.

To estimate the site level covariance matrix we thought of two options. The first option was to use the PES and the census data. The big disadvantage with this approach is that the PES sample is spread over large geographic areas, leading to an underestimate of the site level covariance matrix. The second option was to model the covariances based on the census counts correlations. For design purposes this option is preferred since it is very likely this method overestimates the site level covariances.

For the i-th and j-th poststrata the covariance was estimated by $Cov(\beta_i, \beta_j) = \rho_{ij}\hat{\sigma}_i\hat{\sigma}_j$, where ρ_{ij} is the census counts block level correlations and σ_j is the standard deviation for poststratum j.

Table 1 shows the data for Oakland. The third column shows the population variance of the adjustment factors. These values were adjusted to estimate the standard errors and CVs shown in column 4 to 9. The estimates are shown for several sample sizes. For instance, for Oakland, an estimate of Non-Black Hispanics with a 3 percent CV can be obtained with a simple random sample without replacement of 100 block clusters.

2. Empirical Simulation

This section describes the methodology used for an empirical simulation designed to approximate the total sampling error for estimates of population size for various population subgroups. Sampling error is introduced by sampling the nonrespondents and for coverage estimation. Adjustment factors and the covariance matrix of the adjustment factors were approximated as described in the previous section. The objective of this simulation was to approximate the total sampling error and the relative contribution of each source to the total.

Let $\hat{P}_b$ be the total population estimate (unstratified NRFU block sample),

$$\hat{P}_b = \sum_{b \in S_{ICM}} \beta C_b + \beta \sum_{b \in S_b} W_b C_b + \sum_{b \in Site} C_{b,MR}$$

W_b - NRFU block weight (1-in-6 sample)

$\hat{\beta}$ - adjustment factor defined as before

$C_{b,MR}$ is the population enumerated in mail return questionnaires,

S_{ICM} and S_b are the ICM and NRFU samples respectively.

Note than only the first two terms of the equation contribute to the variance of the population estimate.

To approximate the total sampling error of $\hat{P}_b$ it is easiest to condition on the ICM sample and use the decomposition

$$Var(\hat{P}_b) \doteq \underset{S_{ICM}}{E}\,[\underset{S_b}{Var}(\hat{P}_b|S_{ICM})] + \underset{S_{ICM}}{Var}[\underset{S_b}{E}(\hat{P}_b|S_{ICM})]$$

a.) The second term in the right side of the equation is $E^2(\hat{C})Var(\beta)$, the ICM sampling error.

b.) The first term is $E(\beta^2)Var(\hat{C})$, the NRFU sampling error.

To simulate errors a.) and b.) the following steps were implemented.

i.) Generate 100 β's for each poststrata (SectionIII.C.1). We assumed $\beta_j \sim N(\beta_j, \sigma_j^2)$.

ii.) Draw 100 ICM samples. The samples were drawn independently. A stratified systematic sampling design was assumed.

iii.) For each ICM sample, 100 NRFU samples from the non ICM sample blocks were drawn.

For each NRFU sample estimates of redistricting type data were calculated. For each ICM sample, 100 NRFU samples were selected. To approximate the ICM sampling error component, we first took the expected value of the census estimates across all NRFU samples for each ICM sample and secondly calculated the variance across all ICM samples. The NRFU sampling error component was approximated in a similar way. The effects of undersampling large blocks was also simulated. This simulation process assume a simple random sample of blocks for NRFU. The actual sample design for NRFU is a stratified design with strata similar to those used for ICM sampling. Results from this simulation are available upon request.

IV. Results and Recommendations

In general, both simulations results showed that an ICM sample size in the order of 10 to 15 percent of blocks and a 1-in-6 NRFU sample is sufficient to achieve a coefficient of variation of 1 percent for the estimate of total population. Based on this sample size we will be able to detect a difference between the ICM and census estimates for total population and most population subgroups. In other words, we will be able to assess the effectiveness of the CensusPlus methodology for coverage estimation. Based on these results sample sizes were recommended as follows.

Site	Sample Size	Approximate Housing unit
Oakland	150	10000
Paterson	100	6600
NW LA	100	4600

References

1. Wright, Tommy, "Census Plus: A Sampling and Prediction Approach for the 2000 Census of the United States", August 5, 1993.

2. Schindler, Eric and Navarro, Alfredo, "Census Plus: An Alternative Coverage Methodology" Paper Presented at the 1994 ASA Conference.

3. Woltman, Henry F., "ICM Sample Size Analytical Simulations for the 1995 Test Sites", Internal Census Bureau Memorandum, March 29, 1994.

4. Navarro, Alfredo and Schindler, Eric, "Sample

Sizes for ICM", Internal Census Bureau Memorandum, March 14, 1994.

5. Thompson, John H., "1995 Test - ICM Stratification Technical Documentation", Internal Census Bureau Memorandum, September 7, 1995, DSSD #A-10.

6. Causey, B., Discussion on DSSD #A-2, Unpublished, April 6, 1994.

Acknowledgement

The authors are grateful to Eric Schindler and Lawrence Bates (Bureau of the Census) for expert computational work and many and fruitful statistical discussions prior and during the implementation of the computer simulations.

Footnotes

1┘ The "enrichment factor" indicates what percent of sampling stratum is of a given race. For example, a 50 percent enrichment factor for the Black stratum, indicates that 50 percent of the population in the stratum is Black.

2┘ Blocks were grouped together to form clusters with at least 30 housing units.

Table 1. Summary of Analytical Results
Oakland Summary

Poststratum	Adjustment Factor	S^2_i	CU (%) and SE					
Non Hispanic, White and Other			100		200		300	
			879	.8	607	.6	484	.5
Owner	1.003	7.7						
Renter	1.003	7.7						
Black			900	.5	622	.4	495	.3
Owner	1.078	6.7						
Renter	1.064	9.4						
Hispanic			1447	3.0	1000	2.1	796	1.7
Owner	1.039	9.5						
Renter	1.073	30.7						
Asiand & Pac. Islander			454	.8	314	.6	250	.5
Owner								
Renter	1.005	1.4						
	1.016	3.4						