A COMPARISON OF ALTERNATIVE SAMPLING METHODOLOGIES FOR CENSUS 2000

James Farber and Alfredo Navarro, U.S. Bureau of the Census[1]
James Farber, Decennial Statistical Studies Division, Bureau of the Census, Washington, DC 20233

Key Words: Nonresponse Follow-up, Integrated Coverage Measurement, Undercoverage

Abstract

Since 1970, when the decennial census was first conducted largely by mail, response rates have been declining and undercoverage errors have been increasing. To remedy these problems, the Census Bureau plans to use two major sampling operations in Census 2000. Sampling for nonresponse follow-up will allow the Census Bureau to complete the initial phase in a cost-effective manner, while sampling for Integrated Coverage Measurement will provide an increase in the quality of census data by correcting for coverage errors. Sampling will enable the Census Bureau to achieve the goals of a faster, less costly, and more accurate census.

However, before these sampling techniques can be accepted for use in Census 2000, their potential effect on providing an accurate accounting of the population must be assessed. One way to assess the potential effectiveness of sampling is to compare the errors introduced by sampling to the undercoverage errors of the 1990 census. This paper describes the methodology and results of research into the levels and sources of error from simulations of the sampling operations planned for Census 2000. Additionally, a comparison is made between the sampling errors obtained from these simulations and the undercoverage errors of the 1990 census. This comparison will allow the Census Bureau to determine the optimal enumeration strategy for Census 2000.

I. Introduction

According to the Post-Enumeration Survey (PES), the 1990 Census succeeded in counting 98.4% of the population of the United States. While this level of accuracy is comparable to previous censuses, it was achieved at a significantly greater cost. Of the $2.6 billion census budget, nearly 20 percent was consumed by a single operation: nonresponse follow-up (National Research Council, 1994). Nonresponse follow-up (NRFU) was implemented to count those who did not voluntarily respond in the initial phase of the enumeration. During NRFU, hundreds of thousands of interviewers personally visited all of the nearly 30 million addresses, corresponding to roughly 35% of the mailback universe, from which the Census Bureau did not receive a form. Enumerators resolved many cases in a single visit, but many addresses required several visits before adequate data were gathered. These very hard-to-count addresses cost as much as eighteen times more than a mail return to enumerate (Bureau of the Census, 1996). The high cost of visiting these addresses, combined with their unexpectedly high frequency, greatly contributed to the expense of the NRFU operation.

Despite the large amount of resources devoted to NRFU and other operations designed to count the population correctly, the 1990 Census did not count everyone. This is not too surprising, though, since the population of the United States is too dynamic to count perfectly. However, the census did not simply miss a random subset of the true population. The 1990 PES showed that certain demographic groups were more likely to be missed in the census than other groups. The missed persons tended to be minorities and renters in urban areas, meaning that these groups were disproportionately under-represented in the final census tally. The 1990 PES yielded a set of adjustment factors to correct for undercoverage, but a decision was made not to adjust the census results.

The experience of the 1990 Census clarified the goals for Census 2000: to contain costs and to improve data quality. To meet these seemingly contradictory goals, the Census Bureau plans to use two sampling methods in Census 2000. Sampling for NRFU will reduce the cost of conducting the initial phase. Sampling for Integrated Coverage Measurement (ICM) will reduce the undercoverage bias that has afflicted previous censuses. In this paper we describe these sampling operations, and we also present results of research to compare sampling to 1990 Census procedures to determine if sampling will improve the overall results of the census.

II. Sampling in Census 2000

NRFU SAMPLING

[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.

The Census Bureau has a number of options with respect to NRFU sampling, but three options stand out as the most promising from operational and statistical viewpoints.

The planned NRFU sampling operation will follow what has been termed Direct Sampling. Under this option, the NRFU sample will be selected directly after a period of mail response. All units that have not mailed in a census form during the mailback period will comprise the NRFU sampling universe. A sample will be taken independently within each tract, with the sample size large enough to raise each tract response rate to 90 percent. For tracts in which the initial mail response rate was at least 90 percent, a 1-in-10 sample of addresses will be taken. The addresses selected for the sample will be visited by enumerators for data collection. A statistical model will be used to impute data for nonsampled nonrespondent addresses using data from the sampled nonrespondents and from neighboring mail return units.

A second alternative for NRFU sampling is 90 Percent Truncation. Under this plan, full NRFU will occur after the mail response period. Enumerators will visit as many addresses as required to raise each tract response rate to 90 percent. A 1-in-10 sample of the remaining addresses will then be selected for enumeration, with imputation of data for the nonsampled addresses. A 1-in-10 sample will also be taken in tracts that have a mail response rate of at least 90 percent.

The third alternative is Time Truncation NRFU sampling. Following the mail response period, an intensive outreach effort will be conducted in which enumerators will visit all nonresponding addresses at least once to collect data. After this effort, the NRFU sample will be selected from the addresses that did not complete a census form either by mail or by personal interview. The sample will be large enough to raise each tract response rate to 90 percent. A 1-in-10 sample of addresses will be selected in tracts where the initial response rate is at least 90 percent.

For all three alternatives, the sample will be chosen randomly to ensure that all nonresponding addresses have an equal probability of selection. This also ensures that NRFU sampling will in no way affect coverage - the hardest-to-count housing units will not be sampled at lower or higher rates than other units.

ICM SAMPLING

Since the random NRFU sample will neither improve nor reduce coverage bias, the Census Bureau plans to use ICM sampling to achieve the goal of improved accuracy. ICM will use the same re-interview and comparison method as the 1990 PES, but on a much larger scale. The ICM sample will include 750,000 housing units, compared to 150,000 units in the PES. This large sample size will enable the Census Bureau to produce direct state estimates, which were not possible with the smaller sample of the PES. That is, population and demographic estimates for a state will be based only on sample data collected from housing units in the state. As in the PES, interviews will be conducted at housing units selected for the ICM sample. The data collected from these housing units will be compared to initial phase data from the same units to determine the numbers of persons missed or double-counted in the initial phase of the enumeration. These results will then be generalized to the entire nation using post-stratification. The final census numbers will be produced by combining the results of ICM with the initial phase, increasing the population estimates for post-strata in which many persons were missed and reducing the estimates where persons were double-counted. ICM will thus yield accurate population estimates that will not include the significant undercoverage bias that has persisted through many previous decennial censuses.

III. Sampling Research

The use of sampling in the census means that the final census results will be population estimates instead of counts. Furthermore, these estimates will be subject to sampling error. Since the primary goal of the census is to provide accurate measures of the population, the Census Bureau must assess the potential amount of sampling error to determine if sampling will eliminate more error (undercoverage bias) than it will introduce (sampling variation). In our research, we have reproduced the 1990 Census under simulated Census 2000 conditions to answer this question. We first simulated NRFU sampling in every tract in the nation to obtain NRFU variance estimates. We then simulated ICM sampling using 1990 PES results, and combined the ICM and NRFU variances to obtain a measure of total sampling error. The combined sampling error may be compared to the 1990 undercounts to show which census-taking methodology performs better at the tract level. We also aggregated our results to make comparisons for 103rd congressional districts, states, and the nation.

SIMULATION OF NRFU SAMPLING

Although Direct Sampling is currently planned for use in Census 2000, we implemented all three NRFU sampling alternatives. To be as realistic as possible, we identified the sampling universe for each alternative using 1990 Census operational data that provided such information as how and when each census form was received. That is, for every housing unit in the nation, we knew whether the unit was a mail-return or not, and when each form was

received by census field offices. This information allowed us to use precise definitions of the NRFU sampling universes.

For Direct Sampling, the sampling universe in each tract consists of all occupied, non-mail-return housing units in the tract. For 90 Percent Truncation, the universe is the last 10 percent of occupied, non-mail-return housing units checked in, since the initial follow-up included in this plan means that all but these units will be respondents by the time sampling is implemented. In tracts with a mail response rate above 90 percent, the sampling universe is simply the remaining occupied nonrespondent housing units. For Time Truncation, the NRFU universe is the last 60 percent of occupied, non-mail-return housing units checked in for each tract. We made the assumption, based on previous research, that 40 percent of the units that had been nonrespondents in 1990 would become respondents in Census 2000 due to the outreach effort in this plan. We assumed that this 40 percent will be the first 40 percent of nonrespondents checked in during the 1990 Census, since these units were most likely easier to enumerate than other nonrespondents.

For all NRFU sampling alternatives, we assumed that the U.S. Postal Service will perfectly identify all vacant and non-existent addresses and will return their forms as undeliverable as addressed, and therefore the sampling universes will contain only occupied units. The 1995 Census Test has shown that this assumption will be violated, but we did not have sufficient information to treat these units differently.

SIMULATION OF ICM SAMPLING

To simulate ICM, we used results from the 1990 PES since the PES is the most recent national coverage survey. The 1990 PES yielded a set of adjustment factors and a covariance matrix for 357 post-strata, formed by the cross-classification of census region (groups of states), urban/rural area, race, Hispanic origin, tenure (rent/own status), age, and sex. We assumed that the adjustment factors and covariances will be similar in Census 2000 ICM. Given the fivefold increase in sample size for ICM, this assumption may seem overly simplistic. However, we believe that since ICM will provide direct estimates for each state, the amount of sample information used to obtain those state-level estimates will be approximately equal to the corresponding region-level sample used in the PES. Therefore, the levels of error should also be similar. We made the additional assumption that the adjustment factors and covariances may be used for synthetic estimation at the level of census tracts. This assumption will not hold, but since we made the same assumption in computing tract-level undercounts, we feel that the overall effect on the comparison of sampling errors and undercounts is negligible.

POPULATION AND ERROR ESTIMATION

For each tract, the estimate of total population is:

$$\hat{Y} = \sum_{i=1}^{357} A\hat{F}_i \hat{C}_i$$

$$= \sum_{i=1}^{357} A\hat{F}_i (C_{i,R} + C_{i,Samp} + \hat{C}_{i,NonSamp}) \quad \text{where}$$

$\hat{Y}$ is the final census population estimate for the tract;

$A\hat{F}_i$ is the estimated adjustment factor in post-stratum i;

$\hat{C}_i$ is the initial phase population estimate;

$C_{i,R}$ is the initial phase count in housing units that respond by mail or by personal interview during an initial follow-up;

$C_{i,Samp}$ is the initial phase count in housing units selected for the NRFU sample;

$\hat{C}_{i,NonSamp}$ is the initial phase population estimate for nonrespondent housing units not in the NRFU sample.

Since the NRFU sample is selected randomly, the initial phase population estimate will be unbiased. Therefore, we made the assumption that this estimate is equal to the 1990 Census count for each tract/post-stratum combination. This assumption simplifies variance estimation since the 1990 Census count has no variance. The ICM variance for each tract is then:

$$Var_{ICM}(\hat{Y}) = Var(\sum_{i=1}^{357} A\hat{F}_i C_i)$$

$$= \sum_{i=1}^{357} \sum_{j=1}^{357} C_i C_j Cov(A\hat{F}_i, A\hat{F}_j) \quad \text{where}$$

C_i is the 1990 Census count in post-stratum i.

The NRFU sample is a random sample of housing units from each tract sampling universe, so the variance due to NRFU sampling is:

$$Var_{NRFU}(\hat{C}_i) = (TE-1)S^2 NHU_U \quad \text{where}$$

TE is the Take-Every (the inverse of the sampling rate)

for the tract;

NHU_U is the number of housing units in the tract NRFU sampling universe;

$S^2 = \frac{\sum_{i=1}^{NHU_u}(NP_i - \bar{NP})^2}{NHU_u - 1}$ is the variance of the household sizes of the housing units in the tract NRFU sampling universe (NP_i is the number of persons in the ith housing unit in the sampling universe, and $\bar{NP} = \frac{\sum_{i=1}^{NHU_U} NP_i}{NHU_U}$ is the mean number of persons per housing unit in the tract sampling universe).

The ICM and NRFU sampling operations are completely independent, and therefore their variances may be combined simply by adding them together for each tract. The combined coefficient of variation (CV) for a tract is then

$$CV = \left(\frac{\sqrt{Var_{ICM} + Var_{NRFU}}}{\hat{Y}}\right) \times 100\% \quad .$$

The tract-level undercount rate is simply the relative difference between the 1990 PES population estimate and the 1990 Census count for the tract:

$$UCT = \left(1 - \frac{C}{\hat{Y}}\right) \times 100\% \quad .$$

IV. Results

SAMPLING ERRORS VERSUS UNDERCOUNTS

Table 1 on the last page contains summary statistics for the tract sampling errors and undercounts for total population and for five race/ethnicity groups. It is important to note that the number of tracts has been trimmed to eliminate outliers. There are 60,128 non-vacant tracts in the nation. We first removed tracts that contain only institutional group quarters population, since neither ICM nor NRFU sampling will occur in institutional group quarters. We then removed tracts that have a 1990 Census count of 10 or fewer persons. Relative sampling errors and undercount rates are usually exceptionally large for such small tracts, and do not provide good information for our comparison. Finally, we removed the top and bottom three percent of remaining tracts in the sampling error or undercount distribution. For the race/ethnic groups, we also eliminated tracts that did not have at least one person of the race/ethnic group. By eliminating outlier tracts, we ensure that the comparisons are fair and meaningful. Also, the means shown in the table are weighted by the proportion of total or race/ethnic group population in each tract.

Comparing the sampling alternatives with the undercount, there is no consistent pattern. On average, all of the sampling alternatives do at least as well as the undercount for total population, but this is not true for the race/ethnic groups. Sampling provides small CV's for large groups, such as Non-Hispanic Whites, which are well-represented in nearly all tracts. But sampling does not do as well for groups that are small parts of the population in many tracts. For example, Asians and Pacific Islanders have many tracts with large sampling errors, since there are few tracts with large enough Asian and Pacific Islander populations to support accurate sampling.

However, the means for these groups demonstrate that sampling performs well when accurate estimation is highly important. When a tract has an unusually large population of race group, the type of situation in which poor performance would be most unacceptable, sampling yields an accurate estimate, as shown by the means being much smaller than the maximums for many race groups. The weighted means are pulled down by the small sampling errors in tracts in which a group like Asian and Pacific Islander is the largest part of the population. Thus although sampling can clearly perform poorly for race/ethnic groups in some tracts, it yields accurate estimates in tracts where lack of accuracy would lead to significant change in the final census population estimates.

For congressional districts, states, and the nation, sampling performs better than 1990 Census procedures, as shown in Table 2 on the last page. For the 435 congressional districts, the maximum sampling error is 1.1%, which is lower than the weighted mean undercount rate. This difference is even larger for states, where the maximum sampling error is 0.5% since the ICM sample is designed to yield either a 0.5% CV or a standard error of 60,000 persons, whichever provides the minimum error. The national CV is 0.1%, more than an order of magnitude smaller than the national undercount rate. The sampling errors are identical for all three NRFU sampling plans since nearly all of the error comes from ICM at these high geographic levels.

COMPARING THE NRFU SAMPLING PLANS

Clearly, Direct Sampling is the optimal choice for NRFU

sampling from a statistical viewpoint. It consistently provides the lowest mean error, the lowest maximum value, and the tightest range of errors. Operationally, Direct Sampling is also the best choice of the three options. Direct Sampling entails only one operation: collection of data after selection of the NRFU sample. Both of the other alternatives include two stages of data collection. Under 90 Percent Truncation, complete NRFU is conducted until the 90 percent response threshold is reached in each tract, followed by enumeration of only the sampled addresses. This plan requires daily monitoring of response rates for each of more than 60,000 tracts. Time Truncation includes complete NRFU until each nonresponding address has been visited, after which enumeration occurs at the sampled addresses. The field requirements of these plans increase their costs and complexity relative to Direct Sampling. Since Direct Sampling is both statistically and operationally more attractive, it is the best alternative for NRFU sampling in Census 2000.

SOURCES OF ERROR

At low geographic levels, such as census tracts, the population is usually small and uniform enough that ICM variation is limited. Since the adjustment factor covariance matrix is identical for any geographic level, population size largely determines the ICM variance, as shown by the formula above. Tracts are fairly small geographic units, averaging approximately 4000 persons, so tracts will have much smaller ICM variances than congressional districts or states. And since the NRFU sample is often fairly small, NRFU variation is also limited for census tracts. On average, the tract-level combined variance is split almost equally between NRFU and ICM. This means, however, that Direct Sampling, which always has the largest NRFU sampling universe, may have tract sampling errors twice as large as other plans for small groups, such as American Indians, that may not be represented in other plans' sampling universes.

For higher geographic areas, the ICM variance dominates the NRFU variance since the ICM sample becomes much larger than the NRFU sample for larger areas. At the level of congressional districts, NRFU variation is only about five percent of the total variation, and for states the NRFU contribution is nearly zero. This means that the NRFU sampling plan used is a much more important factor for small area estimation than for estimation of large areas, where the design of ICM is more influential.

V. Conclusions

Our research has shown that sampling will generally perform better than 1990 Census procedures for census tracts, congressional districts, states, and the nation. The difference is more striking for larger geographic areas. For tracts, though, sampling may perform worse for small tracts. For estimation of race/ethnic group populations, sampling is highly accurate for large groups, such as Non-Hispanic Whites, but can falter for smaller groups. However, in tracts where a group is a large part of the population, sampling does well, and therefore the overall effect on the population estimates is positive.

These results, combined with the resource savings of sampling, mean that the Census Bureau should continue with its plan to use sampling methods in Census 2000. The population estimates will be more accurate, and will be obtained using less time and money than in the 1990 Census.

VI. References

Bureau of the Census (1996), "The Plan for Census 200," internal memorandum, Washington, D.C.

National Research Council (1994), *Counting People in the Information Age*, Washington: National Academy Press.

Table 1. Comparison of Alternatives for Total Population and Five Race/Ethnic Groups for Census Tracts

Race/Ethnicity	Sampling Error Range (Weighted Mean)			1990 Weighted Mean Undercount Rate	Number of Tracts
	Direct Sampling	90 Percent Truncation	Time Truncation		
All Persons	0.6%-2.7% (1.1%)	0.8%-3.5% (1.6%)	0.7%-2.7% (1.3%)	1.6%	55,933
Non-Hispanic White	0.7%-7.2% (1.1%)	0.8%-9.0% (1.6%)	0.7%-6.7% (1.3%)	0.9%	55,902
African American	0.9%-61.5% (3.0%)	0.8%-80.9% (5.0%)	0.8%-63.8% (3.6%)	4.5%	53,465
Asian and Pacific Islander	1.2%-63.7% (6.1%)	1.2%-81.9% (9.7%)	1.2%-65.3% (7.5%)	2.6%	53,150
American Indian/Alaska Native	0.5%-73.8% (9.2%)	0.5%-85.0% (9.2%)	0.5%-67.4% (8.1%)	4.6%	52,727
Hispanic American	1.3%-54.6% (3.8%)	1.2%-74.2% (5.7%)	1.2%-56.6% (4.4%)	5.1%	55,386

Table 2. Comparison of Alternatives for Total Population for Congressional Districts, States, and the Nation

Geographic Area	Sampling Error Range (Mean) for ICM and Any NRFU Sampling Plan	1990 Weighted Mean Undercount Rate
Congressional Districts	0.4% - 1.1% (0.6%)	1.6%
States	0.2% - 0.5% (0.5%)	1.6%
Nation	0.1%	1.6%