DISCUSSION: SAMPLING FROM THE 1995 CENSUS TEST BUFFET

Alan M. Zaslavsky, Harvard University
Department of Statistics, 1 Oxford Street, Cambridge, MA 02139

Key Words: Integrated Coverage Measurement, Nonresponse Followup

Many of us at these meetings in Disney enclave in Orlando have looked hard each night for an affordable and interesting meal. Last night I went to dinner at the Coral Cove Cafe on 1995 Test Census Theme Night. I found this out when a waitress approached me holding a CAPI machine and asked a question about my total available assets. Nonetheless, I sampled the buffet, and I would like to tell you about my meal. (Because it was a buffet, the order in which I ate the dishes was slightly different from that in which they are listed on the menu.) Incidentally, one problem with eating when I did (for which the cooks are not responsible) is that I came before the 1995 results were fully cooked, so some of the dishes were actually prepared with leftovers from the last sitting (the 1990 census and Post Enumeration Survey).

Main Course: Petroni, Ikeda, and Singh on the overall design of sampling in the test

This dish was my main course because it included the most complete presentation of the design and goals of sampling in the Test Census, as well as discussing many interesting research questions arising from the new design. I would like to stress the differences between sampling for nonresponse followup (NRFU) and for the Integrated Coverage Measurement (ICM) survey, which are implicit in this paper. The NRFU sample must support estimates of the distribution of a large part (all mail nonrespondents, possibly 30%) of the population. For this large a group, small-area detail is quite important. Fortunately, the sample is likely to be relatively large: perhaps 10–30% of the nonrespondents, or 3–9% of the total population. There is also a lot of auxiliary information, namely the address list which tells us the location of the nonresponding housing units. The ICM survey, conversely, measures a relatively small part of the population (net undercoverage in 1990 of under 1.5%, or gross errors somewhat larger than that), detecting systematic biases over large domains using a relatively small sample (probably around 750,000 households or about 0.7% in 2000). Evidently, both must be successful to make the census work: the NRFU sample is required for small-area detail, and ICM is required to correct remaining biases (differential coverage).

The new design has important implications for the census process. First, the tight schedule of successive operations (mailout questionnaire, NRFU, ICM survey) means that late returns cannot be added to the primary roster as they were in the 1980 and 1990 censuses (or at least they could not be in 1995). Last-minute "coverage improvement" programs will not be feasible under this schedule; omission of these programs should also yield significant cost savings. Late returns could be used as a source of imputations for estimated households, however, while maintaining consistency with estimates sample-based estimates.

Search-match operations (looking for duplicate records in nearby blocks) change under the new design, because with sample NRFU there no longer be will full rosters of actually-enumerated households with identifying information. This will affect estimation of the number of people who could have been enumerated in two different places. The authors discuss alternative methodologies, which include estimation and imputation, for search-match and processing of late returns, under the new circumstances.

Finally, households will have to be imputed to represent those estimated using NRFU and ICM data. These should be drawn from the pool of similar households, i.e. nonrespondents in the same area. Methods for accomplishing this are the subject of lively research, e.g. Zanutto and Zaslavsky (1995).

The key *statistical* issues from the Test have nothing to do with site-wide estimates, although it is important to show that it is *operationally* possible to create these estimates. There are two comparisons that will be watched carefully: (1) characteristics of estimated nonrespondent households under unit vs. block NRFU sampling, and (2) differences between ICM sample and nonsample blocks on indicators of the census process such as late returns. The first tells us whether it is too difficult to find households scattered about in a unit sample, and the second tells us whether the ICM survey has an effect on the census itself.

Bread and Butter: Navarro and Woltman on ICM sample design

Sample design is a staple of creating a new survey. The objectives and conditions for sample design for the Test are quite different, however, from those for the actual census. As noted above, the main statistical questions for the test involve certain subgroup comparisons, rather than estimation for the entire test site. Furthermore, each test site is relatively homogeneous and therefore it is not surprising that Navarro and Woltman's simulations show small gains for disproportionate sample allocation within site; this conclusion cannot be extrapolated directly to a national survey.

Another unsurprising conclusion is that variances for the Census-Plus estimator and dual system estimator (DSE) are very similar. In fact, the *estimators* themselves differ only in the handling of the small "fourth cell," under equivalent *data-collection procedures.*

This course, which at first looked at first sight like a simple piece of white bread, turned out to be a cleverly baked puff pastry, because the lack of complete data from a small area comparable to the test sites made it necessary to synthesize a simulation population using geographically scattered and demographically diverse blocks from the 1990 Post Enumeration Survey (PES). The authors had to apply complex adjustments to make these populations comparable to those in the sites. A different strategy was adopted by the authors of the next paper.

Salad Bar: Town and Fay on variance estimation

A nutritious survey meal must include a variance estimation salad. Variance estimates were controversial in 1990 because with the empirical Bayes small-area estimators that were used then (and that may also have a role in 2000), both point estimates and posterior variances are sensitive to estimates of sampling variance. Furthermore, variance estimates from the Test will be important in tests of the key hypotheses mentioned above.

This was no simple dish of iceberg lettuce, but a rich plate full of olives, ham and cheese. The authors' elegant formulation begins with an unified description of the overall sampling scheme for NRFU and ICM as a form of double sampling, with a corresponding combined estimator. A resampling methodology (modified jackknife) provides a variance estimate. I endorse the use of resampling variance estimates in this context because complex estimators may be used for nonrespondents who are not in the NRFU sample, making analytical estimators for variances are difficult or impossible to obtain.

However, this salad may be more elegant than needed for the 2000 dinner. In that setting, the ICM sample will be very sparse compared to the NRFU sample (much more than in the test), so the two sampling processes may be regarded as almost independent. In that case, it may be adequate (and much simpler) to calculate sampling error from NRFU and ICM separately. Furthermore, in 2000 the domains for NRFU estimation will be much smaller than the estimation cells for ICM, so a unified estimator could be complicated. Nonetheless, it will still be critical to calculate good variance estimates in 2000. For domains smaller than those for which ICM yields direct estimates, these will include variance components for ICM sampling, lack of fit for the model that brings ICM estimates down to smaller domains, and NRFU sampling; the problems involved in estimating and combining these three components will be somewhat different than those confronted in the Test.

Because of the complex structure of the populations, we need honest simulation evaluations of our variance estimators. By an "honest evaluation" I mean one that includes the following steps:

1. Create or find a full (known) population similar to the real population we will be estimating;
2. Simulate repeated sampling from this population;
3. Calculate both the estimator and the variance estimator for each sample;
4. Evaluate whether the expectation of the variance estimator is close to the observed sampling variance.

As in many such evaluations, the first step is the most challenging. The authors create a population by a complex match of Test Census site blocks to blocks in the 1990 PES, which loses the original geographic structure and requires extensive replication of some blocks. It is difficult to tell whether the simulated population that is obtained is similar enough to the true population to give an entirely convincing simulation (although it is still honest in the sense given above). Simulations of estimators for NRFU sampling alone can be relatively unproblematic, because we have comparable populations in the 1990 census (which had no NRFU sampling); simulated populations for ICM sampling (or for PES sampling, in the 1990 evaluations) are more difficult to come by.

Finally, for salad dressing, I pose a question to the authors: why is the variance larger for systematic sampling than for stratified sampling? This phenomenon typically occurs when there is a pattern in the data that matches the skip interval of systematic sampling. I see no reason why this should be the case and suspect that the discrepancy is an artifact of the way the population was synthesized or (more likely) an accident of sampling.

Drinks: Mulry and Navarro on evaluation of sampled data- collection procedures and estimators

Such a rich dinner must be washed down with some evaluations (non-alcoholic, because this is federally supported research), which anticipate and address some of the most likely criticisms of the data-collection procedures involved in NRFU and ICM sampling.

The key comparison for NRFU sampling is between coverage under a block and unit NRFU sample. This comparison is critical because the ICM sample will be a *block* sample (due to the infeasibility of controlling an independent survey in dispersed households). Although a *unit* NRFU sample may yield smaller variances, if its coverage properties are different from those of the block sample, we have no way of measuring them.

Mulry and Navarro propose a variety of outcome and process measures for this comparison of NRFU results under the two sample designs, including mean household size, vacancy rates, and add and delete rates in NRFU. (They also consider the distribution of household sizes, but a χ^2 test with many categories has low power and is difficult to interpret, so I would suggest replacing it with more focused contrasts.) An important purpose of variance estimation in the Test Census is to obtain valid tests of differences on these measures.

The main issue concerning ICM is the comparison between the Census-Plus (C+) estimator and dual system estimation (DSE). Only the estimators, but not alternative field procedures, are compared in the tests.

Some of the comparisons are predictable. Because C+ and DSE differ primarily in that the latter includes an estimate of the unobserved "fourth cell," we can expect that DSE estimates of population will exceed those from C+, and that the difference will be largest in the areas with most severe undercoverage.

Because ICM represents the final effort to estimate the coverage properties of all preceding stages of estimation, ultimately the acceptability of ICM estimates depends on the face validity of the patterns of coverage that it detects. I regard the use of Wilcoxon tests for relationships between block undercoverage rates and predictors of undercoverage as somewhat arbitrary. Nonparametric tests lend an appearance of objectivity which does not correspond to the actual properties of the test in this setting. Because the number of coverage errors in any block is usually small, rank tests will be very affected by the noise in the block-level data. I would prefer to emphasize tests that smooth or average patterns across blocks (as the final estimator does), such as Pearson correlations or ANOVA and regression models. These measures have the additional advantage that the analysis can use covariates at the household as well as the block level.

Dessert: Thibaudeau and Navarro on NRFU sampling rates

This was a particular enjoyable dessert because it satisfied a craving I have felt for a systematic analysis of optimal sample allocation in NRFU sampling. Thibaudeau and Navarro's work represents a big advance over previous *ad hoc* proposals.

The authors take a systematic approach to optimal design, which includes the following steps: (1) define a simplified mathematical model that summarizes the main features of the problem, (2) state the objectives and constraints of the design, (3) optimize the design parameters (here, sampling rates), (4) look at the practical implications of the solution, and (5) modify the solution to consider other goals and constraints not included in the original formulation.

The empirical results, based on analysis of 1990 census data, are quite impressive and suggest substantial improvements over *ad hoc* alternatives, such as doing 100% enumeration until a common target completion rate is attained everywhere and then completing the enumeration at a fixed sampling rate. The figure illustrates this comparison, showing the total NRFU workload as a function of mailback response rate (assuming all other characteristics constant, and with some arbitrary choices of the parameters of the two plans) for the authors' rule (curve) and the *ad hoc* plan (bent line). The graph suggests that the latter is not entirely misguided, but nonetheless the optimal design is substantially better.

The authors directly tackle the constraint of what we might call an "equitable distribution of accuracy," which they formulate as equal coefficients of variation for local areas of roughly equal population. This same issue is also important for ICM, where it is complicated by the fact that an important unit of analysis for ICM, the states, comprises domains of widely varying populations. (See an article by Kadane and discussions by Mulry and Zaslavsky, forthcoming in *Journal of Official Statistics,* for more on this problem.) Note that as long as estimates for local areas are approximately unbiased, accuracy for large areas does not depend much on the optimality of the sample design for small areas. The cost of poor local accuracy is that local uses, such as districting and planning, will suffer; from this point of view, inaccuracy imposes costs on the local area even if its population estimate is correct in expectation.

The following are some other considerations in this problem, of which the first few have already been considered by the authors. Operationally, it may be necessary to simplify the range of sampling rates (by using a discrete set of rates). It may also be politically necessary to have a floor on the sampling rates. There may be a large return on an initial period of nonsampled NRFU targeting the easy-to-count cases; if true, this might affect the authors' conclusion that it is suboptimal to have any amount of nonsampled NRFU. Predictability and even distribution of the NRFU workload may be operationally important; this consideration supports the authors' rejection of nonsampled NRFU. Other objectives could be defined, with slightly different design implications (for example, Jabine for example, minimization of *average* rather than *maximum* squared coefficient of variation would shift sample slightly toward higher-response areas, where better accuracy would then be obtained.

Conclusion

In conclusion, these papers constituted an excellent menu for the 1995 Census Test Buffet. We now look forward to the 2000 Census *Prix Fixe* Dinner!

Jabine, T.B. (1976), "Equity in the Allocation of Funds Based on Sample Data," *Statistical Issues in Allocating Federal Funds and Estimation of Local Government Finances,* Washington: Bureau of the Census, 2–8.

Zanutto, E.M. and Zaslavsky, A.M. (1995), "A Model for Imputing Nonsample Households with Sampled Nonresponse Follow-up," *Proceedings of the Section on Survey Research Methods, American Statistical Association.*

Percent of population sampled as a function of mailback response rate under two sample allocation schemes.