OVERVIEW OF THE U.S. CENSUS 2000 LONG FORM DIRECT VARIANCE ESTIMATION

Philip M. Gbur and Lisa D. Fairchild, U.S. Bureau of the Census
Philip M. Gbur, 7975 Central Park Circle, Alexandria, VA 22309

Key Words: Systematic Sample, Successive Differences

I. Introduction

A systematic sample of addresses in Census 2000 received a long form, or sample, questionnaire which collects detailed socioeconomic and demographic characteristics. The selected addresses and the people at those addresses were weighted to represent the entire population and housing stock of the nation. Variances were estimated for a subset of resulting long form estimates using a successive difference replication (SDR) methodology. (See [2] for a description of the SDR methodology.) Extensive research was done prior to the 1980 Census on alternative variance estimators [1] and we selected the SDR methodology based on these results and experience with the SDR on the Current Population Survey and other projects within the Census Bureau.

However, the long form sample can be the basis of a myriad of estimates calculated at many geographic levels. The Census Bureau has a commitment to provide estimates of sampling error for all estimates and to minimize burden on data users by not overwhelming them with volumes of error estimates. Thus, we will provide a set of design factors to approximate sampling errors. The following sections present a brief overview of the sample design and the weighting and a description of the direct variance estimation of the long form questionnaire data for the Census 2000. We will also describe the components which were changed from initial plans and from 1990.

II. Sample Design

The addresses that were to receive the long form questionnaire were chosen by taking a systematic, variable rate sample of addresses. The ultimate goal was to sample roughly 17 percent of all addresses nationwide. This was achieved through appropriate application of the selected sampling rates to each governmental unit or census tract. Application of the rates for Census 2000 was based on the interim census tract delineation, as updated census tracts were not yet available. Governmental units were defined as states, counties, cities, incorporated places, school districts, American Indian Reservations, Tribal Jurisdiction Statistical Areas (now known as Oklahoma Tribal Statistical Areas), minor civil divisions is selected states, and census designated places in Hawaii.

The rates used were: 1-in-2, 1-in-4, 1-in-6 and 1-in-8, and were applied based on a governmental unit's or tract's predetermined measure of size. The number of occupied housing units was used as the measure of size.

The sampling rates were applied at the block level. For blocks that fell into more than one sampling stratum, we applied the higher sampling rate.

The sampling strata and their cutoff points were:
- 1-in-2 for governmental units < 800 housing units;
- 1-in-4 for governmental units between 800 and 1200 housing units;

if not 1-in-2 or 1-in-4; then
- 1-in-6 for census tracts < 2000 housing units; and
- 1-in-8 for census tracts $\geq$ 2000 housing units.

The following rates were used for certain data collections and special populations:

a. Group Quarters were sampled at a 1-in-6 rate.

b. Service Sites (such as shelters and soup kitchens) were sampled at a 1-in-6 rate.

c. The Telephone Questionnaire Assistance (TQA) operation took incoming calls for requests for mailing questionnaires and for interviews. Interviews were done for short forms only and individuals providing an interview were not eligible for long form sampling. Individuals who telephoned to request a questionnaire received either their designated form type or were subject to a 1-in-6 sampling rate, depending upon whether they had their census identification number.

d. Addresses added to the mailout universe after the initial sampling were sampled according to the sampling rate of the stratum that the addresses' block was in.

III. Weighting

As in every census since 1940, when we introduced

This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review than official Census Bureau publications. This report is released to inform interested parties of research and to encourage discussion.

content sampling, the iterative proportional fitting methodology was used in the Census 2000 to estimate the characteristics of the entire country based on the long form sample. We carry out this methodology, also known as raking, within weighting areas.

Weighting areas, the geographic level at which we conduct the weighting, were formed within counties. Weighting areas were required to have a minimum of 400 sampled persons. As necessary, small counties with fewer than the prescribed number of cases were allowed to stand alone as weighting areas.

To ensure that we have a basic minimal sample within the weighting areas, there was augmentation of the long form sample using a set of predetermined rules, as needed. This was done to attain a minimum observed sampling rate within each area, reducing the associated variance. Long form data were imputed from short forms for sample augmentation. Augmentation of sample counts used the smallest number of addresses needed to reach the desired minimum observed sampling rate within each weighting area. After augmentation, weighting proceeded separately for people, occupied housing units, and vacant housing units.

For each sample unit we set an initial weight equal to the inverse of the observed sampling rate (100 percent count divided by the number of sample cases received). We then carried out the iterative proportional fitting methodology, also known as raking. Raking was performed in several stages.

For person weighting, for each weighting area, we formed a four-dimensional matrix using household type and size (such as family with own children with four people and family without own children with four people), sampling rate, whether the person is a householder, and Hispanic origin by race and age/sex. For occupied housing units, we used three dimensions: household type by size; race and Hispanic origin of the householder by tenure; and sampling rate. Vacant housing units were weighted based on a three cell array: "for sale;" "for rent;" and "other." If a given classification/cell was not sufficiently large, then it was collapsed with another classification following a predefined pattern.

Raking is an iterative proportional adjustment of the cross-classified cell counts. The interior cell counts within a classification were multiplied by the ratio of the 100 percent count (for that classification) to the initially inflated sample total (for that classification). An iteration of the raking consists of one stage of adjustment for each dimension. Each stage adjusts all interior cell counts by the appropriate cell ratio. The raking progressed until a predefined stopping criterion was reached.

The final step in the weighting process was to integerize the post-raking weights using a controlled rounding procedure.

Further details on the Census 2000 weighting process may be found in [4].

IV. Direct Variance Design

A. Census 2000 Methodology

For Census 2000, we used the SDR methodology to calculate the direct variances at the weighting area level. The SDR methodology has several expected advantages which caused us to select it for use. Primarily, it better reflects the systematic nature of the sampling. In addition, it has been researched extensively and is currently being used for the American Community Survey and the Current Population Survey at the U.S. Census Bureau.

The SDR methodology was developed by Robert Fay, based on the successive difference variance estimator for systematic samples. A successive differences estimator calculates the variance from the sum of squares of differences from overlapping pairs of sample units. This allows order of selection to be taken into account, when the units' order of selection is maintained within the calculation.

For example, say a 1-in-6 sample of $n = 4$ housing units is selected in order $j = 1, 2, 3, 4$ and it is recorded whether or not each unit is owner occupied ($x_j = 1, 1, 0, 0$, where owner occupied = 1 and not owner occupied = 0). The estimated owner occupied total, $\hat{X}$, is the sum of the weighted values for the units, $\sum_{j=1}^{n} w_j x_j$ = 6(1) + 6(1) + 6(0) + 6(0) = 6 + 6 = 12 housing units. The successive difference variance is

$$Var(\hat{X}) = (1-f)\frac{n}{2(n-1)}\sum_{j=2}^{n}(w_j x_j - w_{j-1}x_{j-1})^2,$$

where f is the sampling fraction. So,

$$Var(\hat{X}) = (1-\frac{1}{6})\frac{4}{2(4-1)}\left((6-6)^2 + (0-6)^2 + (0-0)^2\right)$$

$$= 20$$

or the standard error is 4.5 housing units.

For Census 2000, designated and observed sampling

units were identified from various data files. Sample units were assigned overlapping pairs of row numbers from a Hadamard matrix of order 52. (A Hadamard matrix of order J is a J x J matrix with elements -1 or +1 only and orthogonal columns. See [7] for further description of Hadamard matrices.) Hadamard matrix row numbers were assigned to the sample units, designated or enumerated on a long form, according to their order of selection within each independent sample. The sample unit types were housing units, group quarters (GQ) persons, and service based enumeration (SBE) persons. Persons within households received the same row assignments as their housing unit. Order of selection was not available for GQ or SBE persons (who were not augmented), but because each GQ or SBE site was independently sampled, we did not expect the order of selection within site to be informative. An arbitrary sort order was set within site for row assignment purposes.

The Hadamard matrix values were matched to the sample units for each of 52 replicates. The row assignments were the matrix row numbers and the replicate number was the matrix column number. Each sample unit was assigned the corresponding matrix values for each replicate.

For observed sample units, replicate factor values were calculated incorporating a finite population correction factor. The replicate factor calculation was based on the assigned Hadamard matrix values and is defined as:

$$f_{ir} = 1 + [(2)^{-\frac{3}{2}} a_{i1,r} - (2)^{-\frac{3}{2}} a_{i2,r}](1 - f_o)^{1/2}$$

Where:

- f_{ir} is the replicate factor for the i[th] sample unit and the r[th] replicate; $i = 1, \ldots n$; $r = 1, \ldots, 52$;
- $a_{i+1,r}$, $a_{i+2,r}$ is a value (+1 or -1) from a Hadamard matrix of order 52 which corresponds to the i+1[th] or i+2[th] row and r[th] column for the i[th] sample unit; and
- f_o is the observed sampling rate in the weighting area.

The replicate weights were calculated using the replicate factors and the integerized final weights from the weighting process. Fifty-two replicate weights were calculated for every housing unit and person in the observed long form sample.

Standard errors were calculated separately for characteristics of persons, families, and housing units/households. Replicate factors were multiplied by the final weights to produce replicate final weights. Once replicate final weights were produced, the SDR method estimates the standard error, S_{SDRt}, of the estimator for the t^{th} data item through the formula:

$$S_{SDR,t} = \left(\frac{4}{52}\sum_{r=1}^{52} [x_{rt}(\frac{X_c}{X_{rc}}) - x_{ot}]^2\right)^{\frac{1}{2}}$$

Where:

- x_{rt} is the weighted total of the r[th] replicate for data item t where $r = 1, \ldots, 52$;
- X_c is the weighted total of the sample for data characteristic c where characteristics could be persons, housing units, or families;
- X_{rc} is the weighted total of the r[th] replicate for data characteristic c; and
- x_{ot} is the weighted total of the sample for data item t.

In addition, standard errors were calculated assuming a 1-in-6 simple random sample (S_{SRS}) for each, t[th], data item as follows:

If $x_{ot} \leq 0.98\ X_c$, then

$$S_{SRS,t} = (5x_{ot}[1 - (x_{ot}/X_c)])^{1/2} ;$$

else if $x_{ot} > 0.98\ X_c$, then

$$S_{SRS,t} = (5x_{ot}[1 - 0.98])^{1/2} .$$

The design factor (DF) was calculated from these standard errors for the t[th] data item at the weighting area level as:

$$DF = S_{SDR,t} / S_{SRS,t} .$$

B. Modifications from Our Initial Plan

The SDR methodology was implemented for Census 2000 with minimal advance testing of the methodology or of the processing in a decennial environment. Thus, we knew that, as with the implementation of any new approach, there may be risks. As a result, it was not surprising that modifications were made from our initial plan before reaching the design described above. The following sections detail the primary modifications made and the supporting rationale.

1. Number of replicates and reweighting

We had initially planned on using 100 replicates and reweighting the replicate estimates from replicate initial weights to the same census control totals used in the production long form weighting. For each weighting

area, the post-collapsing structure of the matrices from the production process would have been used to reweight each replicate. This would have allowed the variance estimate to reflect any additional variance (expected to be relatively small) resulting from the weighting process.

However, during early testing, it was determined that we would not be able to implement the fully planned methodology due to computer space and processing limitations. It was decided to reduce the number of replicates to 52 and to eliminate the reweighting to save on computer space and processing time.

2. Estimates close to the total

Generalized variances based on initial runs of the direct variances were higher than in 1990 than could be reasonably explained for some estimates. It was determined that this occurred for estimates which were very close to the total population, for which one would expect a small standard error. Further investigation suggested that we were picking up something close to the estimated variance of the total rather than the variance of the small omitted portion. To compensate, the factor X_c/X_{rc}, was included in the S_{SDR} formula.

Note that had the reweighting been implemented, that would have controlled the replicate estimates to these three characteristic totals (population, housing units, and families), and more, and this factor would not have been necessary. Any group that was controlled to in the raking process after all collapsing would still be controlled to and have a standard error equal to zero.

3. Response adjustment factor

Imputation for missing data introduces a component of variance. We had originally planned, for the first time, to account for item imputation in the long form variances which generally uses a hot deck imputation methodology. A sample size adjustment was to be used [6]. This adjustment implies that item nonresponse increases the estimated variance by the ratio of the full sample size over the nonallocated response count, under certain assumptions. We expected that the variance estimates calculated with the sample size adjustment would be conservative but closer to the "true" variance than variance estimates calculated without it.

The adjustment factor was derived to take into account that S_{SDR} includes the finite population correction factor. See [5] for details of the derivation. The response adjustment factor, RAF_t, was derived as:

$$RAF_t = \sqrt{1+2(n_t / r_t - 1)/(1- f_o)}$$

Where:

n_t is the total count of sample units eligible to respond to data item t in the weighting area;

r_t is the count of sample units with a nonallocated response in the universe for data item t; and

f_o is the observed sampling rate in the weighting area.

We would have expected the sample size adjustment to have improved the variance estimate by reflecting nonresponse, but would give conservative estimates of the variance of imputation. Less conservative ways of estimating the variance of nonresponse could be developed but would have been even more operationally complex. The operational simplicity of the sample size adjustment gave it the best chance for implementation in a production process, but even this approach could not be reliably implemented given time and resources available.

C. Alternatives Considered

Three primary options were initially considered for long form direct variance estimation. They were: (1) a random groups (RG) variance estimator, as used in 1990; (2) a Jackknife (JK) variance estimator; and (3) the SDR approach. The SDR estimator would be carried out in a similar manner as described above. We describe the JK and RG methods below. See [7] for further details on these methodologies.

The JK estimator is based on the sum of the squared differences of pseudo-subsample estimates from the average of these pseudo-subsample estimates. Initially, g subsamples are systematically selected from the full sample. The i^{th} pseudo-subsample is composed of the g-1 subsamples left when the i^{th} subsample is left out. Thus, g pseudo-subsamples are created.

Each pseudo-subsample is independently reweighted and then the weighted totals are formed. The variance is found as:

$$Var(\hat{X}) = (1-f_o)\,\frac{g}{g-1}\sum_{j=1}^{g}\left(\hat{X}_j - \sum_{j=1}^{g}\frac{\hat{X}_j}{g}\right)^2 .$$

Where:

f_o is the sampling fraction;

g is the number of pseudo-samples (studied in [1], g = 4, 8, and 12); and

$\hat{X}_j$ is the weighted total for the jth pseudo-subsample.

The RG methodology was used in the 1990 Census for estimating long form direct variances. The procedure for the random groups estimator starts with systematically subdividing the weighting area samples into *g* subsamples. For the 1990 Census, *g* was set to 25. The calculation of the estimated variance for a particular estimate may then proceed through the formula:

$$Var(\hat{X}) = (1 - f_o)\ \frac{25}{24} \sum_{i=1}^{25} (\hat{X}_i - \frac{\hat{X}}{25})^2 \ .$$

Where:

f_o is the observed sampling rate in the weighting area;

$\hat{X}_i$ is the weighted total of the characteristic in a weighting area based on the records assigned to the i^{th} subsample; and

$\hat{X}$ is the sum of the 25 values of $\hat{X}_i$ -- that is, $\hat{X} = \sum_{i=1}^{25} \hat{X}_i$.

D. Changes from 1990

The random groups methodology was implemented for the 1990 Census long form variances and, as described above, we used the SDR methodology for the Census 2000 direct variances. The SDR methodology takes the order of selection into account which the random groups methodology does not. Also, the SRS standard error was allowed to go to zero in 1990 for estimates equal to or close to the total. The design factor methodology was used in 1990 for the generalized variances.

V. Generalized Variances

The generalized variance methodology is similar to that used for the 1990 census. It begins with the calculation of design factors. Design factors are the ratio of the standard error, S_{SDR}, from the direct variance estimate for the complex design over the standard error estimate, S_{SRS}, assuming a 1-in-6 simple random sample as described above.

DFs were calculated for selected data items within each weighting area. Due to space limitations, generalized design factors will be made available across four percent-in-sample categories or intervals. The percent-in-sample was defined at the weighting area level to be the percent observed unweighted sample count out of the 100 percent count, which was equal to the final weighting area observed sampling rate multiplied by 100. The count was of persons for population characteristics and of housing units for housing characteristics.

Data items were arranged into groups and subgroups based on characteristic. For each state, the District of Columbia, and Puerto Rico, generalized design factors for each group and subgroup were calculated over each of the percent-in-sample intervals as a weighted average design factor. They were also calculated at the national level.

Data item groups were examined for homogeneity of variance. Specific data item design factors which were determined to be outliers were down weighted.

VI. Future Research

Future work may be done in developing these procedures for other surveys such as the American Community Survey. Work may be done on accounting for item nonresponse, specifically, on how to define item nonresponse for a variable that is a combination of two or more component items. In addition, further effort may be devoted to identifying efficiencies in implementation of raking and reweighting. This could lead to an increase in their feasability as a production system.

ACKNOWLEDGEMENTS

The authors wish to thank all of the many people whose efforts supported the production of the direct variances for Census 2000. The authors also thank Raj Singh for his guidance and input and Sam Hawala and Elizabeth Huang for their helpful comments on drafts of this paper.

REFERENCES

[1] Fan, Milton C., et. al., "1980 Census Variance Estimation Procedure," Proceedings of the Survey Research Methods Section of the American Statistical Association, pp. 176-181, 1981.

[2] Fay, Robert E. and George F. Train, "Aspects of Survey and Model-Based Postcensal Estimation of Income and Poverty Characteristics for States and Counties," Proceedings of the Government Statistics Section of the American Statistical Association, pp. 154-159, 1995.

[3] Hansen, Morris H., William N. Hurwitz, and William G. Madow (1953), Sample Survey Methods and Theory Volume 1 Methods and Applications, John Wiley & Sons, Inc., New York, New York, p. 469.

[4] Hefter, Steven, and Gbur, Philip M.(2002), "Overview of the U.S. Census 2000 Long Form

Weighting," 2002 Proceedings of the American Statistical Association, Survey Research Methods Section [CD-ROM], Alexandria, VA: American Statistical Association, to appear.
[5] Kim, Jae-Kwang and Michael Brick, "Sample Size Adjustment Method for Long Form Imputation Variance Estimation, internal Westat memo, July 3, 2001.
[6] Little, Roderick J. A. and Donald B. Rubin (1987), Statistical Analysis with Missing Data, New York: John Wiley & Sons.
[7] Wolter, Kirk M. (1985), Introduction to Variance Estimation, New York: Springer-Verlag.