VARIANCE ESTIMATION IN THE CENSUS 2000 A.C.E. SUBSAMPLES

**Douglas Olson, Decennial Statistical Studies Division,
U.S. Census Bureau, Washington DC, 20233-7613**

Keywords: Replication, Jackknife Variance Estimation

Abstract

Stratified random sampling at differential rates is commonly used in social science surveys to improve the chances of getting sufficiently large samples of sub-populations of interest. Multi-phase sampling is often used to further target desired populations. The variances of such samples can be estimated using jackknifing, which creates replicated values of the characteristic of interest as if one sampling unit had not been part of the sample, then sums the squared differences between those replicates and the population total. Kim (2000) suggested a method for estimating the variance in a three-phase design in which the first two phases were stratified and the third was a simple random sample from among the sample units in the second phase. In this paper, Kim's method is extended to a three-phase design in which all three phases are stratified samples at differential rates, but the last phase is stratified by characteristics not related to those used in stratifying the first two phases.

Introduction

The Evaluation Follow-up Interview (EFU) and Measurement Error Reinterview (MER) studied the assignment of Enumeration status (Correct, Erroneous and Unresolved) in the Enumeration sample (E sample) of the Census 2000 Accuracy and Coverage Evaluation (A.C.E.). Persons included in the study were assigned a revised enumeration status based on information obtained from a reinterview and rematching. Because the values in this review were derived from sampling, the weighted totals have a variance that must be estimated. The standard errors originally published with the EFU review were calculated using a drop-one stratified jackknife, the stratification being the one used to draw the EFU sample from the A.C.E. sample. This paper attempts to estimate variances using a stratified jackknife methodology that respects each of the three main phases of sampling and appropriately weights each replicated value to respect the covariances implicit in the sample design.

Sampling Methodology

The EFU sample was a subsample taken from the two-phase A.C.E. sample of block clusters. A block cluster is a group of one or more blocks, which are geographic areas similar to city blocks, although instead of necessarily being bounded by roads, the boundaries could be railroad tracks, bodies of water or non-physical political boundaries. The block cluster was the primary sampling unit for the design described here. The results to be presented will reflect the characteristics of persons who reside within those clusters.

The first phase of A.C.E. sampling assigned every block cluster in the United States into a stratum defined by:

- the state in which it was located
- its size, (small, medium or large) as measured by the number of housing units it contained, or whether it was on an American Indian Reservation (AIR)

Hence, there were 204 theoretically possible first-phase sampling strata possible from crossing 50 states and the District of Columbia with three sizes or AIR location, although only 179 contained any block clusters. There were 29,136 clusters containing housing units selected into the first phase of the sample, selected from among all the clusters in the country.

The second phase of the A.C.E. sample partitioned the strata of the first phase according to:

- the minority population of the cluster
- the consistency between two independent counts of housing units

There were 11,303 block clusters selected into the second-phase sample. This was the principal sample of block clusters used in the Census 2000 A.C.E. (ZuWallack, et al 2000)

The EFU sample, the third phase, was drawn from the clusters of the A.C.E. sample for evaluation purposes. It was systematically selected at differential rates from among the clusters in the second-phase sample, stratified using a set of characteristics relevant to the

evaluation, but without regard to the strata definitions used in the first two phases:

- Geographic Region (Northeast, Southeast, Midwest, West)
- Minority population density (Hispanic, Other Minority, Non-Minority)
- "Problem" status, defined by characteristics of particular interest in evaluation

There were 24 theoretically possible EFU strata (4 Regions x 3 Minority groups x 2 Problem statuses), although the Midwest Hispanic and Other Minority definitions were collapsed together reducing the number of strata to 22. The clusters in these strata were sampled at different rates by the "interestingness" of the cluster for evaluation purposes.

- Problem clusters were included in EFU with certainty (425 of 425)
- Non-Problem Hispanic or Minority clusters at 1-in-4 (705 of 2,825)
- Non-Problem Non-minority clusters at 1-in-7.283 (1,102 of 8,026)
- 27 clusters had no relevant cases for the EFU and were excluded

At each phase of sampling, cluster weights were assigned as the reciprocal of the sampling probability within the phase, so that by the end of the third phase, each cluster had a weight equal to the product of the three phase-weights, termed hereafter the "EFU Weight". (Keathley 2001)

$$W_{hgi} = N_h / n_h \times n_{hg} / r_{hg} \times N_{e(hgi)} / n_{e(hgi)} \quad (1)$$

$W_{hgi} = Weight\ of\ cluster\ hgi$

$h = First-phase\ sampling\ stratum$

$g = Second-phase\ sampling\ stratum$

$i = Block\ cluster$

$n_{hg} = First-phase\ clusters\ in\ second-phase\ stratum\ hg$

$r_{hg} = Clusters\ selected\ in\ second-phase\ sample$

$e(hgi) = EFU\ stratum\ to\ which\ cluster\ hgi\ belongs$

$N_{e(hgi)} = Second-phase\ clusters\ in\ EFU\ stratum$

The samples have used the block cluster as the primary sampling unit and the methods discussed here estimate the variance associated with block cluster sampling. The use of this sample, however, is to estimate characteristic values for persons who live in the block clusters, assigning each person a weight equal to the block cluster weight. In actual practice, some additional sampling operations and adjustments have been made to the weights of individual persons beyond the cluster weights described here. These adjustments were formed to handle unusual cases and to reduce workload in the largest clusters. The final weight assigned each person in the EFU reflects the effect of those weights, but the variance estimation methodology presented here ignores the covariances implicit in those within-cluster samplings. Experience from previous samples suggests they have only a small effect on the results presented here.

Estimation Methodology

The EFU review compared the assignment of Enumeration status (Correct, Erroneous or Unresolved) of the sample persons as assigned in the A.C.E. and the Evaluation. The results are a 3x3 comparison of the sample-weighted number of persons in each category. Since the tabulation represents counts of persons, the results are the totals of the final EFU Weights. An analogous scheme can be used to create counts of any other person characteristic.

$$Y = \sum_g \sum_h \sum_i \sum_j W_{hgij} X(hgij) \quad (2)$$

$Y = Sample\ Estimate\ of\ population\ total\ of\ characteristic\ Y$

$X(hgij) = Characteristic\ value\ of\ Y\ for\ person\ hgij$

In the EFU, Y is the pair of enumeration statuses from the A.C.E. and the Evaluation; and X is an indicator that person $hgij$ belongs to the group defined by Y.

Variance Estimation Methodology

The general method to calculate a drop-one jackknife sample is to remove each primary sampling unit (in our case the block cluster) and calculate what effect its removal would have on the variable of interest. The

sum of the squares of those removal effects is the estimate of the variance:

$$\widehat{Var}(Y) = \sum_{k}^{K} c_k (\hat{Y}^{(k)} - \hat{Y})^2 \qquad (3)$$

k is the index of each primary sampling unit

c_k is a factor associated with replicate k

$\hat{Y}^{(k)}$ is the replicate value of $\hat{Y}$

The replicate values $\hat{Y}^{(k)}$ are estimates of what $\hat{Y}$ would be if sampling unit k had not been included in the sample and the rest of the sample was re-weighted to reflect k's exclusion. The factor c_k is usually slightly less than 1, reflection the slightly smaller sample size used in $Y^{(k)}$. The replicate value is calculated by re-weighting the contribution to *Y* of each of the sampling units (block clusters) by a factor $\alpha_i^{(k)}$, whose calculation will be explained later in this section:

$$\hat{Y}^{(k)} = \sum_{i} \alpha_i^{(k)} Y_i \qquad (4)$$

Y_i is the weighted characteristic value of Y in block cluster i

The published counts of the EFU Review were accompanied by standard error estimates calculated using a drop-one stratified jackknife, stratified on the EFU strata only (Krejsa, 2001). Hence it is a conditional variance, conditioned on the particular A.C.E. (i.e. second-phase) sample.

Table I
Results of the Evaluation Follow-up to the Census 2000 A.C.E.[1]

Population Estimate (Standard Error)	Evaluation Coding		
Production Coding	Correct	Unresolved	Erroneous
Correct	247,114,898 (6,337,607)	1,424,770 (254,488)	2,827,414 (223,469)
Unre-solved	2,873,110 (400,351)	3,010,280 (203,352)	928,719 (117,602)
Erroneous	908,385 (99,380)	124,641 (23,369)	3,118,191 (202,575)

[1]Source: Krejsda and Raglin (2001)

Because the EFU study was an evaluation of Census procedures, it was appropriate to estimate the variance conditionally. But if it was deemed desirable to estimate characteristics of the U.S. population using the EFU sample, the accompanying variances would need to reflect the sampling that went into selecting the A.C.E. sample.

Kim et. al. (2000) designed a replication strategy for a three-phase stratified jackknife in which some A.C.E. clusters (second phase) had been sampled for inclusion in a field operation called Targeted Extended Search (TES). The TES sample included 1,089 clusters from a universe of 5,326 that contained persons eligible to benefit from TES operations. It was a systematic random sample drawn with equal probabilities, without regard to the A.C.E. sample design. The EFU sample, our third phase, resembles the TES sample in disregarding the strata from the first two phases, but differs in that the third phase is another stratified sample at differential rates, while the TES was a simple random sample from a single universe.

Following the method of Kott (1997), Kim (2000) had designed a set of reweighting factors to apply to the second-phase A.C.E. weights when calculating replicate values. These weights adjust for the first two phases of sampling:

$$\alpha_{hgi}^{[stu]} = \begin{cases} 0 & hgi = stu \\ \frac{r_{hg}}{r_{hg}-1}\frac{n_{hg}-1}{n_{hg}}\frac{n_h}{n_h-1} & hg = st, u \neq i, I(stu)=1 \\ \frac{n_{hg}-1}{n_{hg}}\frac{n_h}{n_h-1} & hg = st, I(stu)=0 \\ \frac{n_h}{n_h-1} & h = s, g \neq t \\ 1 & h \neq s \end{cases} \quad (5)$$

$I(stu)$ indicates if stu included in second phase sample

stu = Block whose associated replicate is being calculated

hgi = Another block cluster whose contribution to replicate stu is calculated

u = index of cluster within second phase stratum st within first phase stratum s

i = index of cluster within second phase stratum hg within first phase stratum h

Note: It might aid comprehension to think of "stu" and "hgi" analogously. 's' and 'h' are first-phase stratum designations, 't' and 'g' are second-phase, and 'i' and 'j' are individual clusters within poststrata "st" and "hg."

These factors were applied to the weighted totals of each cluster in calculating the replicate totals. Note that replicates are calculated for dropping out the clusters sampled out by the second phase even though they have zero data values.

Adjusting for TES sampling required calculating an additional reweighting factor for the TES sample persons:

$$TES^{stu} = \frac{\sum_h \sum_g \sum_k \alpha_{hgk}^{(stu)} I_{hgk}(\text{in TES universe})}{\sum_h \sum_g \sum_k \alpha_{hgk}^{(stu)} I_{hgk}(\text{in TES sample})} \quad (6)$$

So the final reweighting factor of cluster ghi for replicate stu became:

$$R_{hgi}^{(stu)} = \alpha_{hgi}^{(stu)} x TES^{stu}$$

The same technique can be adapted for use in calculating replicate reweighting factors for the EFU sample. An additional weight to be applied to each cluster *hgi* when calculating its weight for inclusion in replicate *stu* is:

$$EFU^{stu} = \frac{\sum_R \sum_h \sum_i \alpha_{hgi}^{[stu]} I_{hgi}(e_{stu}) I_{hgi}(\text{in EFU sample})}{\sum_R \sum_h \sum_i \alpha_{hgi}^{[stu]} I_{hgi}(e_{stu}) I_{hgi}(\text{in EFU universe})} \quad (7)$$

$I_{hgi}(e_{stu})$ = indicator that hgk in same EFU stratum as stu

Note that clusters *hgk* and *stu* can only be in the same EFU stratum if both are part of the second phase sample.

This factor is necessary because the EFU sample was drawn from among clusters selected into the A.C.E. at different rates. The different probabilities of inclusion in the A.C.E. affects the probabilities that the other clusters would be selected. Application of this factor to adjust for EFU sampling creates the final weighting scheme for jackknifing:

$$R_{hgi}^{stu} = \begin{cases} \alpha_{hgi}^{stu} x EFU^{stu} & \text{if } ghi, stu \text{ in same EFU stratum} \\ \alpha_{hgi}^{stu} & \text{otherwise} \end{cases} \quad (8)$$

This α_{hgi}^{stu} is the $\alpha_i^{(k)}$ in equation (4), with stu=(k) and hgi the "i" in that equation.

Results

Table II: Results
Standard Error of the Evaluation Follow-up using three-phase Variance Estimation

	Evaluation Coding		
Production Coding	Correct	Unresolved	Erroneous
Correct	7,775,280	287,583	232,097
Unresolved	406,515	217,148	121,916
Erroneous	104,836	23,641	211,443

Most of the three-phase standard errors in the EFU Review were similar to those estimated using the one-phase stratum jackknife of the published total but are all greater, which should be the case because the three-phase design reflects the contribution to variance of the first two phases while the one-phase design does not.

The most important difference is in the Correct/Correct cell, whose standard error is 23% larger than the one-phase error, while the others are all 1-7% larger. This probably reflects the fact that the Correct/Correct cell includes 94% of the overall sample, so its variance is driven primarily by the variance of the overall sample size, while the others are driven primarily by their percent of prevalence within the sample.

Discussion

As of this writing, another evaluation is underway to study the effect of further operations similar to EFU using the same sample but slightly different methods. The jackknife used above could be adapted for use in dual system estimation, in which both systems would be similar to EFU.

Limitations

Several samples and other operations that contribute to variance have been ignored or simplified for ease of programming, believed to have only very minor influence on results:

- Missing Data – Imputations based on sample totals introduce variance into the assignment of persons into different categories. The variance associated with that imputation is ignored here.
- Collapsing of second-phase strata that included only one cluster – A.C.E. variance method collapsed these clusters into a new second-phase stratum that was not part of the first-phase stratum, but maintained the original first-phase stratum in computation; new method keeps the second-phase stratum from the A.C.E. but assigns it to the first-phase stratum from which the new second-phase stratum was a subset.
- TES sampling – Covariances between cluster selection probabilities used in TES sampling were ignored.
- Within-cluster subsampling – Subsampling within clusters was used to reduce workload in the A.C.E. (in large block clusters) and in EFU (among the least interesting evaluation cases). The person weights used in this paper reflect all phases of subsampling, but no additional effort has been made to estimate variances associated with those samples.

Caveat

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

References

Kim, Jae Kwang; Navarro, Alfredo and Fuller, Wayne (2000). Variance Estimation for the 2000 Census Coverage Estimates. *Proceedings of the JSM*, 2000, pp. 515-520.

Krejsa, Elizabeth A. and Raglin, David A. (2001). ESCAPII: Evaluation Results for Changes in A.C.E. Enumeration Status. ESCAP II Report #2, U.S. Census Bureau web site

Kott, P.S. and Stukel, D.M. (1997). Can the Jackknife be used with a Two-phase Sample? *Survey Methodology*, 23, pp. 81-89.

Zuwallack, Randal; Salganik, Matthew; Cromar, Ryan and Mule Jr., Vincent Thomas (2000). Final Sample Design for the Census 2000 Accuracy and Coverage Evalution. *Proceedings of the JSM*, 2000, pp. 441-446.

Keathley, Don H.(2000). EFU Sample Design, Stratification, Selection, and Weighting. Planning, Research, and Evaluation Divison TXE/2010 Memorandum Series: CM-GES-W-02