Simplified Variance Estimation for Complex Surveys

Eric Schindler[1], Bureau of the Census, Washington, DC 20233

Keywords: Simple Jackknife, Stratified Jackknife

ABSTRACT: Many surveys employ complex sampling and estimation designs. Complex designs lead to complex variance estimation systems requiring significant investments of resources over long time periods to develop, implement, and run. The Census 2000 Accuracy and Coverage Evaluation (A.C.E.) Survey with a complex multi-stage design, imputation, and post-stratified dual-system estimation is an example. The A.C.E. design resulted in a complex variance estimation system using the stratified Jackknife approach. The development of this system involved external consultants, several years' effort for several persons, and an intense double programming effort. A simplified variance system was also programmed. This system used the final weights, ignored the intermediate stages of sampling, ignored the variance from a low level of imputation, and estimated a simple Jackknife variance. Total design and programming time could be measured in days for one person. Run time was also significantly reduced. The results of the simplified program for estimates of housing unit coverage are compared to those from the full variance estimation system.

BACKGROUND

Coverage in the 2000 United States Decennial Census was measured by the Accuracy and Coverage Evaluation (A.C.E.) program with Capture/Recapture or Dual System Estimation (DSE) methodology. Coverage estimates were made for persons and for housing units. During the evaluations of person estimates, serious problems were discovered in the estimated coverage for persons. It was decided that coverage adjusted person estimates would not be used for standard census releases at least until corrected estimates could be produced. Therefore, the discussion in this paper is limited to the estimated coverage for housing units, although some of the same analysis was performed for the original estimates of person coverage with similar results. The A.C.E. was implemented for a multi-stage sample of about 300,000 housing units in about 10,000 block clusters. Housing units on the Master Address File were grouped into block clusters with about 30 housing units each. Some 30,000 block clusters with 750,000 housing units were sampled from the Master Address File with over 100 million records in 7 million blocks. The selected clusters were sent to the field offices where independent lists of housing units were compiled. A subsample of 10,000 block clusters was selected with oversampling of block clusters when the census housing unit count and the independent listing housing unit count were discrepant. Some housing units which could not be matched within the block cluster were eligible for further subsampling in a targeted extended search (TES) operation. An additional one or two stages of sampling of block clusters and housing units was selected, and the surrounding clusters were also searched for the selected TES eligible housing units.

Estimates of Coverage Correction Factors (CCF) were calculated for 98 post-strata defined by the variables vacant/occupied, race of householder, Census Region, size of metropolitan area/type of enumeration area, size of structure. For each post-stratum **i**, $\mathbf{CCF_i}$ was estimated by: $CCF_i = \frac{CE_i}{E_i} \times \frac{P_i}{M_i}$, where:

- $\mathbf{E_i}$ is the weighted number of housing units in the census in sample block clusters post-stratum **i**. This set of housing units is called the E-sample or enumeration sample. E-sample housing units are examined to determine which are correctly enumerated.
- $\mathbf{CE_i}$ is the weighted number of correctly enumerated housing units in the E-sample in post-stratum **i**.
- $\mathbf{P_i}$ is the weighted number of housing units in the independent sample in sample block clusters post-stratum **i**. This set of housing units is called the P-sample or person sample. P-sample housing units are examined to determine which are really housing units and which of the real housing units can be matched to the census in the block cluster or the TES search area.
- $\mathbf{M_i}$ is the weighted number of housing units in the P-sample in post-stratum **i** which could be matched to housing units in the census.

The dual system estimate for post-stratum **i** is obtained by multiplying the census count of housing units in the post-stratum, $\mathbf{CEN_i}$, times $\mathbf{CCF_i}$.

The correct enumeration or match status could not be determined for a small percentage of housing units. Using a small number of imputation classes, these units are assigned a probability of correct enumeration or of match based on the final status of units with similar initial coding which were able to be resolved.

Synthetic estimates of housing units for geographic or other domains are made by adding the products of the post-stratum CCFs by census count of housing units in the domain and post-stratum. The estimated national

[1]Eric Schindler is a mathematical statistician in the U.S. Census Bureau's Decennial Statistical Studies Division. This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a Census Bureau 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.

undercount rate for all housing units was 0.61 percent. The census count of 115,877,639 housing units was estimated to be 708,819 lower than DSE of 116,586,458.

PRODUCTION ERROR ESTIMATES

The official variance system used in production reflected the imputation of missing correct enumeration or match status and the three or four stages of sampling which were (1) from the universe to 30,000 block clusters, (2) from 30,000 to 10,000 block clusters, and (3) for TES eligible housing units. All of these stages were combined in a stratified Jackknife procedure.

The DSE is a three phase sample employing both double expansion estimators and reweighted expansion estimators. Calculation of 30,000 sets of replicated correct enumeration and match rates for the few unresolved cases was performed within the basic replication of weights for cases not dropped out. Reflecting the multiple stages of sampling resulted in the proliferation of multiple cases within the basic replication formulae. (See Sands and Navarro, 2001 for details.)

The basic design was developed by Dr. Jae Kwang Kim with assistance from Census Bureau experts and Professor Wayne Fuller. Computer specifications were developed by Michael Starsinik. The parallel systems for persons and housing units were programmed, mostly in FORTRAN, by the Census Bureau's systems staff and double programmed, mostly in SAS, by Census Bureau statisticians. Because of the complexities of the sample design and the resulting error estimates, considerable staff resources were required to design and implement the systems. The error estimation systems were separate from the estimation systems, were run after the estimation systems, were allowed an additional several weeks in a tight schedule, and required substantial run time.

The production standard error for the total housing unit DSE of 116,586,458 was 188,121.

SIMPLIFIED ERROR ESTIMATES

Inspiration

The author was assigned the tasks of triple programming the estimation and calculating simplified error estimates intended as a quick check to catch gross errors either in the specifications or the implementation of the production error estimation system. These activities occur after the creation and editing of files of census housing unit counts and files with the A.C.E. sample results. Three basic assumptions were made:

- **The variance from the imputation of correct enumeration or match probabilities is small.** Approximately 0.0021 percent of the census housing units required the imputation of a correct enumeration probability and less than 0.0001 percent of the independent sample required the imputation of a match probability. Since the imputed probabilities were fractional, removing one block cluster in the stratified Jackknife results in minimal change in the estimated probabilities and should have little effect on the error estimates. Even if, as the author would prefer, an actual status were imputed instead of a probability, the proportion of housing units requiring imputation is so small that the effect would be negligible. On the other hand, preliminary research did show that at imputation levels in the 10 percent range, the variation from the imputation process becomes significant and should be reflected in the variance design.
- **The variance from the intermediate stages of sampling is small.** Removing a block cluster which does not appear in the final sample has a small effect on the weights of other block clusters in the same sampling stratum. Since the DSE estimates are really ratio estimates of CCFs which are then multiplied by census counts, a slight change in the weights affects both the numerator and the denominator and should not affect the replicate ratio estimate very much. Later, we will see that this assumption affects estimates of error across Census regions.
- **A simple Jackknife procedure is adequate**. Wolter (1985) discusses simplifications such as those employed here for imputation and reweighting when discussing random groups. Kott and Stukel (1997) show empirically that the Jackknife, under suitable conditions, provides suitable estimates in a two-phase sample when calculating ratios. Our own empirical experience for the similar 1990 Post Enumeration Survey and for the complex A.C.E. design described below leads to the same conclusion for the circumstances which we faced.

There are many advantages of a simpler system. Design time is reduced; standard errors are available as soon as the estimates are available; the potential for errors in specifications or implementation is reduced; and, perhaps equally important for visible programs such as the decennial census, the simple Jackknife procedure can be understood by the non-statisticians who supply the Census Bureau's funds and use the Census Bureau's data.

Statistical Design

The decision was to employ a simple Jackknife using the final in-sample block clusters as the primary sample units. An equally simple alternative would be to use a random group design, but the block cluster design naturally lends itself to the Jackknife.

(Elapsed time : 10 seconds)

Systems Design

This is simple a four step process:

- Write a program to calculate the DSEs, which had to be done anyway for verification purposes.
- Copy the program to the bottom of itself.
- Modify this lower portion of the program to remove one block cluster at a time and redo estimates for each replicate.
- Calculate standard errors.
  (Elapsed time for steps 2, 3, and 4 : 10 minutes)

The design chosen requires that the input files be read twice. An alternative design reading the files only once can be implemented if summary data for each block cluster are stored as the files are read. Both designs are straightforward.

Implement Systems Design

The task is to implement the four steps described above. The author admits to having had the advantage of writing similar programs for several census tests and the Census 2000 Dress Rehearsal and for person DSEs, but this helps mainly with the first step.

- Write program to do estimates. With slightly more complex requirements than for the previous versions for census tests and the Dress Rehearsal this required several input file preparation programs and a completely new main program. Several days of programming effort were required. All estimates of levels and rates match the production estimates exactly. Programs were written in FORTRAN because the author finds that it has more flexibility than and is faster than SAS or other packages for numerical manipulations.
- Copy program to the bottom of itself.
- Read the input E-sample and P-sample files again removing one block cluster at a time from the totals created when calculating the full sample estimates and redo estimates. This is really a simple process of remembering the number of housing unit records in each block cluster with housing units in the final stage of sampling and subtraction. This process produces one set of replicate CCFs for each of the 10,267 block clusters with housing units.
- Calculate standard errors and a variance-covariance matrix using the simple Jackknife formulae. One replication is calculated for each block cluster, **b**, in the final sample with housing units. For each post-stratum **i**, and replicate **b**, the replicated coverage factors is given by:

  $$CCF_{i,b} = \frac{CE_{i,b}}{E_{i,b}} \times \frac{P_{i,b}}{M_{i,b}},$$

  where each term is obtained by subtracting the number of Correct Enumerations or E-sample housing units or P-sample housing units or matched housing units in block cluster b from the corresponding total. There is no reweighting to reflect the stages of sampling. There is no recalculation of imputed correct enumeration or match probabilities. The standard error is given by:

  $$SE(CCF_i) = \sqrt{\frac{n-1}{n}\sum_{1}^{n}(CCF_{i,b} - \overline{CCF_{i,b}})^2} ,$$

  where **n** is the total number of block clusters with housing units. (Programming and debugging time for steps 2, 3, and 4 : 4 hours.)

The error estimates are integrated into the main program so error estimates and a variance-covariance matrix are available immediately. The run time for the main program on a VAX alpha machine including estimates for post-strata, state and national totals, and ten subtotals within each state is less than five minutes. The creation of input files of housing unit counts and the creation of files with the A.C.E. survey results required considerably longer to run but were required for the estimates with or without error estimates.

The main estimation program can be designed to produce synthetic estimates at lower levels of geography either directly or from the variance-covariance matrix.

Variance of Variance

In order to evaluate the simplified error system it is necessary to compare the results with those from the full-scale production system, so it would be nice to have the standard error of the difference between the production and the simplified error estimates. This would be extremely complicated. However, estimating the standard error of the simplified standard error is relatively easy if we simply Jackknife the Jackknife standard error. That is:

- Remove one of the 10,267 block clusters at a time.
- Calculate standard errors for the remaining 10,266 block clusters by the Jackknife procedure described above. This process replicates each post-stratum CCF estimate over one hundred million times and results in 10,267 sets of error estimates, so it takes a long time to run, approximately sixty hours on the VAX alpha machine.
- Calculate the standard errors of the error estimates.

This process took a week on and off to program because the testing run times were longer. To speed up the debugging process ten block clusters at a time were dropped, so "only" 1,027 sets of replicated standard errors were calculated. The replication within each of these 1,027 replicates to calculate the standard error estimates continued to drop one block cluster at a time. This "shortcut" caused many of the bugs which required debugging. Even then, a minimal run required a half-hour or more of run time.

We would have preferred the variance of the difference between the error estimates. Because of the correlation between the two sets of error estimates, the results tell us more about the practical significance of the differences in the standard error estimates than about the statistical significance.

COMPARISON OF RESULTS

Of course, the major question is: How well did the simplified variances do? It was certainly fast and easy, but several components of the variance were omitted, potentially leading to underestimates of the standard errors. Are the results usable? In the case of the Census 2000 housing unit DSE, the answer is a resounding **YES**.

Post-stratum Comparisons

Over the 98 post-strata in the housing unit DSE design, the average percent difference in the estimated standard errors (Simple Jackknife/Production - 1) is about -0.79 percent. It makes sense that the simple Jackknife error estimates are lower, as they are for 88 post-strata, because of the components of the error not accounted for. If the percent differences are weighted by the housing unit counts, the percent difference is about -0.51 percent. Thus, there is little practical difference between the two sets of error estimates.

In most surveys the variance of the variance is high and the Census 2000 A.C.E. estimates for housing units are no exception. The coefficients of variation for the simple Jackknife standard error for post-stratum CCF estimates range from 0.12 to 0.98 with a median at 0.29 and a 90th percentile at about 0.60. However, the small numerical differences between the production and the simple Jackknife standard errors translate into small statistical differences. The differences in the standard errors divided by the standard error of the simple Jackknife standard error, $\frac{SE_{SimpleJK} - SE_{Production}}{SE(SE_{SimpleJK})}$,range from -0.19to 0.06. While it is true that the two sets of error estimates are correlated because they are based on the same sample, it is unlikely that the correlations are so large that the differences are statistically significant.

The estimates of the standard error of the simple Jackknife standard error dropping ten block clusters at a time with 1,027 replicates average about 3 percent larger than those dropping one block cluster at a time. The range is from 40 percent larger to 10 percent smaller.

National Comparisons

Table 1 shows the original count, the DSE, the two standard error estimates and the percent difference for the national total and ten subgroups: (1) vacant units, (2) occupied units, units with a (3) Black or (4) Hispanic or (5) Asian or (6) Native Hawaiian/Pacific Islander householder, (7) units on an American Indian Reservation with an American Indian or Alaska Native householder, (8) units with a White or Some Other Race or American Indian or Alaska Native but not on a reservation householder, (9) single family units, and (10) multi-family

Table 1: National Estimates

	Housing Unit Count	Dual System Estimate	SE_{Prod}	$SE_{SimpleJK}$	% Diff	$SE(SE_{SimpleJK})$ Drop 1	$SE(SE_{SimpleJK})$ Drop 10	$\frac{SE_{SimpleJK} - SE_{Production}}{SE(SE_{SimpleJK})}$
total	115,877,639	116,586,458	188,121	194,274	3.27%	9,229	9,474	0.6668
vacant	10,414,216	10,777,553	109,411	110,417	0.92%	9,471	9,733	0.1062
occup	105,463,423	105,808,904	136,032	139,018	2.19%	7,746	7,826	0.3854
White	80,707,902	81,069,552	113,914	115,468	1.36%	7,392	7,600	0.2103
Black	12,119,949	12,066,685	34,641	34,503	-0.40%	4,831	4,899	-0.0285
Hisp	9,213,313	9,231,299	32,617	32,791	0.53%	2,770	2,980	0.0626
Asian	3,115,320	3,122,088	19,038	18,992	-0.24%	3,437	3,473	-0.0133
NHPI	157,755	167,243	5,006	4,971	-0.69%	1,642	1,703	-0.0210
AIR	149,184	152,038	2,273	2,294	0.96%	312	352	0.0697
single	84,277,111	85,008,508	135,227	138,671	2.55%	7,870	8,005	0.4376
multi	31,600,528	31,577,950	129,223	129,487	0.20%	10,054	10,528	0.0263

units.

Reversing the situation for the individual post-strata, the simple Jackknife standard error for these synthetic estimates is often a few percent larger than the production standard error. As will be seen later, these differences are generally larger than those found for state estimates. The last three columns of Table 1 show the two estimates of the simple Jackknife standard error of the state estimates. The estimate of the difference over the standard error of the standard error for the national total (0.6668) is about three times larger than that observed for any of the state totals or state subtotals. Once again, there is little difference whether we drop one or ten clusters at a time to replicate the error estimates. Once again the differences between the production error estimates and the simple Jackknife error estimates are not practically significant. The high correlations between the production and simplified estimates may mean that some of the differences are statistically significant.

State Level Comparisons

At the state level, similar results we obtained. The simple Jackknife standard estimates are within a few percent of the production standard error estimates for almost all estimates compared. The differences in the standard errors are all smaller than three-tenths of the standard error of the Jackknife standard error. Table 2 shows results for the 545 non-zero estimates for the 51 states for the eleven totals in Table 1.
The simple Jackknife standard error estimates are generally slightly lower than the production system standard errors. However, the differences are acceptable in terms of having little practical significance. The standard error estimates are all within 0.3 standard errors of the simple Jackknife standard error. Barring very high correlations, this means that the differences are not statistically significant either.

Table 2: 545 State Level Estimates

Percent Difference		Difference in CV	
$\frac{SE_{SimpleJK} - SE_{Production}}{SE_{Production}}$		$\frac{SE_{SimpleJK} - SE_{Production}}{SE(SE_{SimpleJK})}$	
-15% to -10%	1		
-10% to -5%	6		
-5% to -2%	4	-0.3 to -0.2	9
-2% to -1%	47	-0.2 to -0.1	3
-1% to 0%	294	-0.1 to 0.0	340
0% to 1%	165	0.0 to 0.1	168
1% to 2%	28	0.1 to 0.2	24
		0.2 to 0.3	1

Variance-Covariance Matrix

The one area where real differences were observed between the production estimates and simple Jackknife estimates was in the variance-covariance matrices. For the large estimates along the diagonals and for related post-strata, the estimates matched fairly closely, with the production estimates often a few percent larger than the simple Jackknife estimates. Off the diagonal, and especially when the two post-strata were for different Census regions, the simple Jackknife estimates were close to 0 because the methodology reflected only the post-stratification. The production estimates, on the other hand, were much larger in magnitude (although still much smaller than near the diagonal) and usually negative. Recall that the production estimates reflected the third and fourth stages of sampling for the targeted extended search. The TES stages crossed Census region boundaries. Given the slight increases at the national level, it does not seem likely that many important estimates would be seriously affected by these differences.

LIMITATIONS

- The simple Jackknife ignores potentially important components of the variance. However, the empirical evidence shows that the omitted components are not practically significant. The national estimates and the variance-covariance matrix show the effects of this assumption, but the effects are manageable.
- Using the standard error of the simple Jackknife standard error does not incorporate the correlation between the two sets of error estimates. The differences between some pairs of error estimates could be statistically significant.
    - Estimating the standard error of the difference of the error estimates would require an enormous statistical and programming effort and perhaps a prohibitive amount of computer time.
    - Since the differences are not practically significant, does it matter if some of them are statistically significant?
- The results shown are for a particular ratio estimator under the conditions observed in the Census 2000 A.C.E. housing unit estimation process. They should not be generalized to cases with much higher rates of imputation or using substantially different estimators. For example, preliminary estimates from the Census long form sample show that simple Jackknife standard errors for estimates of levels instead of ratios are not as close to the estimates from the production system based on successive difference replication (Gbur and Fairchild, 2002). Differences often reach 30 percent.

CONCLUSIONS AND RECOMMENDATIONS

The principal conclusion of this paper is very simple: **there are negligible differences between the error estimates produced by the production system and those produced by the simple Jackknife procedure for the Census 2000 housing unit DSEs.** Similar results were obtained in the original estimates for person DSEs, and similar results are expected for revised person estimates.

The principal recommendation is similarly simple: **easy-to-design, easy-to-implement, easy-to-run, easy-to-explain error estimation methods should be considered for use as the production error estimates; more sophisticated methods should then be used for evaluation.** This is especially so when past experience shows that there is little practical difference between the two alternatives. This approach provides acceptable, easily verified, real time error estimates. At several times in the evaluation of the original person DSEs, the author was asked to provide simple jackknife error estimates because the production system was not yet ready. These were disseminated as *preliminary, subject to change*, but it makes more sense to have the official estimates available at the same time.

The moral of the story paraphrases Deming's *Don't work harder, work smarter.* **Don't work harder, don't work smarter, work simpler.** Consider whether the added layers of sophistication can be expected to achieve anything significant. If not, they belong in the evaluation process with less time pressure and less stringent verification requirements.

REFERENCES

Gbur, Philip M and Lisa **Fairchild** (2002) Overview of the U.S. Census 2000 Direct Variance Estimation, *American Statistical Association 2002 Proceedings, Section on Survey Research Methods*, 2002, to appear

Kott, Phillip S. and Diana M. **Stukel** (1997) Can the Jackknife Be Used With a Two-Phase Sample? *Survey Methodology*, Vol 23, No.2, pp. 81-89.

Sands, Robert D. and Alfredo **Navarro**, (2001) Census Accuracy and Coverage Evaluation Survey Variance Estimates, *American Statistical Association 2001 Proceedings, Section on Survey Research Methods*, 2001

Wolter, Kirk (1985). *Introduction to Variance Estimation*. New York: Springer Verlag. pp.83-85.