

## 2000 Census Accuracy and Coverage Evaluation Survey Variance Estimates

**Robert D. Sands and Alfredo Navarro, Bureau of the Census**

Robert D. Sands, DSSD, Room 2505-2, US Census Bureau, Washington, DC 20233, (301) 457-4255

**Keywords:** Variance estimation, dual system estimates, multi-phase sampling

### Introduction

The dual system estimates (Wolter, 1986, Haines, 2001) produced by the Accuracy and Coverage Evaluation (A.C.E.) Survey are based on random samples and therefore are subject to random sampling errors. The estimates are also subject to non-sampling errors, such as error due to missing data. The A.C.E. sample was designed to produce census coverage estimates at the poststratum level. Estimates of variances and covariances for the dual system estimates (DSEs) were computed through a replication method. This paper presents and examines the estimates of variances and coefficients of variation for the DSEs. A secondary objective is to compare the estimated coefficients of variation (CVs) of the DSEs with the CVs used from the 1990 Post Enumeration Survey (PES) 357 poststrata design (Thompson, 1992).

### Dual System Estimation

The A.C.E. Survey relies on dual system estimation to estimate household population coverage in Census 2000 (Haines, 2001). The Census Bureau obtains a roster of persons from the A.C.E. block clusters independently of the census. The independent person roster (P Sample) and the census person roster (E Sample) are matched; the results of the matching and followup interviewing are used to estimate the total number of persons. Figure 1 illustrates a simplified DSE formula.

**Figure 1.**

		P-Sample		
		In	Out	
E-Sample	In	$X_{11}$	$X_{12}$	$X_{1+}$
	Out	$X_{21}$	$X_{22}$	$X_{2+}$
		$X_{+1}$	$X_{+2}$	$X_{++}$

$$DSE_{++} \equiv X_{++} = X_{1+} \frac{X_{+1}}{X_{11}}$$

where  $X_{..} = \sum_i X_{i..}$ ,  $X_{i..}$  is an indicator variable for person  $i$  falling in a particular cell

$$D\hat{S}E = (C - \Pi) \left( \frac{C\hat{E}}{\hat{N}_e} \right) \left( \frac{\hat{N}_n + \hat{N}_i}{\hat{M}_n + \left( \frac{\hat{M}_o}{\hat{N}_o} \right) \hat{N}_i} \right) \quad (1)$$

where  $C$  is the census count

$\Pi$  is the number of whole person census imputations

$\hat{N}_e$  is the estimated E - sample total

$C\hat{E}$  is the estimated number of correct census enumerations

$\hat{N}_n$  is the estimated P - sample nonmovers

$\hat{N}_o$  is the estimated P - sample outmovers

$\hat{N}_i$  is the estimated P - sample inmovers

$\hat{M}_n$  is the estimated P - sample nonmover matches

$\hat{M}_o$  is the estimated P - sample outmover matches

The production DSE formula (1) is employed in the A.C.E. survey. Referring also to Figure 1, the first two factors on the right side of (1) correspond to the  $X_{1+}$  in Figure 1. This census enumerations term is the actual census count net of the cases that do not have enough information for matching multiplied by the correct enumeration rate of the census. The correct enumeration rate is estimated from A.C.E. operations. Further, the numerator of the remainder of the right side of (1) corresponds to  $X_{+1}$  in Figure 1. This A.C.E. enumerations term is the sum of the estimated number of nonmovers and the estimated number of inmovers. An inmover is a person who moved *into* an A.C.E. sample area between Census Day in April, 2000 and the A.C.E. interview conducted in Summer, 2000. Outmovers and nonmovers are defined analogously. Finally, the denominator of the remainder of the right side of (1) corresponds to  $X_{11}$  in Figure 1. This

matches term is the sum of the estimated nonmover matched persons and the estimated mover matched persons.

Dual System Estimates are calculated separately for population subgroups (poststrata),  $r$ . Poststrata are defined so that persons within a particular poststratum have similar survey capture probabilities. This reduces correlation bias that occurs when the independence assumption of dual system estimation model is violated (Wolter, 1986). Poststratum-level dual system estimates are then used to determine coverage factors which are applied to all people counted in the census according to their specific poststratum (Davis, 2001).

The difference between the original census count and the dual system estimate reflects the coverage of the census, either a net undercount or a net overcount.

### **A.C.E. Sampling Methodology**

The A.C.E. Survey has a three-phase sampling design (Kim, et al, 2000). The first-phase yielded a stratified sample of 29,136 (US) block clusters, the second-phase consisted of a sub-stratification of the 29,136 first-phase block clusters and yielded 11,303 block clusters (bcs) and 300,913 housing units (hus) The second-phase sampling was conducted using four second-phase strata and was conducted independently in each of the four first-phase strata. Finally, the third-phase or Targeted Extended Search (TES) sample consisted of three strata which were defined without regard to the definition of the first- and second-phase strata designations. These third-phase strata were defined as follows: (i) non-TES designated block clusters (not involved in TES sampling or operations), (ii) a non-certainty stratum from which a 10% non-certainty sample of second-phase bcs was selected, and (iii) an additional 10% certainty stratum in which all block clusters were selected for TES operations. The TES sample bcs selected from the (ii) and (iii) third-phase strata then had their surrounding bcs searched for additional matches (Navarro and Olson, 2001). All second-phase sample bcs remained, however, in the final A.C.E. interview sample whether TES designated or not.

The multi-phase A.C.E. Survey design was performed to gain efficiency by using information determined about the sample from an earlier stage. For instance, the second-phase sampling used information, from a housing unit listing operation performed on the 29,136 first-phase sample block clusters, to stratify the first-phase sample. Further, the third-phase sampling used

information, from a housing unit matching operation performed in the 11,303 second-phase sample block clusters, to stratify (designate the type of TES of) the second-phase sample.

It is important to note here that a two-phase stratified sample design is different than a two-stage stratified design. In the latter, the sampling rates for the various strata for each of the stages of sampling is completely predetermined. On the other hand, in two-phase sampling the second-phase sampling rates are determined only after the first-phase sample has been drawn. The second-phase sampling rates are a function of the first-phase sample observations (Rao, 1973; Cochran, 1977).

### **A.C.E. Estimation Methodology**

Considerable theoretical and empirical work has been undertaken to develop proper methods of estimation and replication variance estimation for two (or more)-phase stratified sampling designs (Rao and Shao, 1992; Kott and Stukel, 1997; Kim et al, 2000). To this end, Kott and Stukel (1997) identified two point estimators that are employed in deriving two-phase sample estimates, the Double Expansion Estimator (DEE) and the Reweighted Expansion Estimator (REE). This is important because Kott and Stukel (1997) suggested a consistent jackknife replication variance estimator for the REE only. Kim et al (2000) found, however, that the DEE could be rewritten in an algebraically equivalent form to the REE and, in turn, also produce consistent jackknife replication variance estimates.

The following paragraphs will present how the multi-phase A.C.E. DSE was implemented in the context of the work cited above.

First, the following notation will be used in the remainder of the paper. Let  $A_1$ ,  $A_2$ , and  $A_3$  be the first-, second-, and third-phase samples, respectively. Let the letters  $h$ ,  $g$ , and  $c$  be first-, second-, and third-phase stratum designations, respectively. Let  $w_{hi}$ ,  $\alpha_{hgi}$  and  $\tau_c$  be the first-, second-, and third-phase sampling weights, respectively. Let  $x_{hgi} = 1$  indicate that a block cluster,  $i$ , is in first-phase stratum,  $h$ , and in second-phase stratum,  $g$ . Let  $s_{ci} = 1$  indicate that a block cluster,  $i$ , is in third-phase stratum,  $c$ . Let  $n_h$  be the number of first-phase sample block clusters,  $i$ , in first-phase stratum,  $h$ . Let  $n_{hg}$  be the number of first-phase sample block clusters in the first-phase stratum,  $h$ , and in the second-phase stratum,  $g$ . Let  $r_{hg}$  be the number of second-phase sample block clusters in the combination of strata,  $h$  and  $g$ .

The dual system estimation equation in (1) contains seven different estimated terms. Within any of these seven DSE terms the particular term,  $\hat{y}$ , where  $\hat{y}$  is either  $\hat{N}_e$ ,  $\hat{C}\hat{E}$ ,  $\hat{N}_n$ ,  $\hat{N}_o$ ,  $\hat{N}_i$ ,  $\hat{M}_n$ ,  $\hat{M}_o$ , can be broken into two parts: (i) the total prior to any adjustments made by the third-phase,  $\hat{u}$ , and (ii) the additional contribution from the TES operation following the third-phase sampling,  $\hat{v}$ . The TES part,  $\hat{v}$ , is further broken down into two sub-parts (c=2) the portion due to TES non-certainty sampling and (c=3) the portion due to TES certainty sampling (2).

$$\hat{y} = \sum_{i \in A_2} \alpha_{hgi} \hat{u}_i + \sum_{c=2}^3 \sum_{i \in A_3} \alpha_{hgi} \tau_c s_{ci} \hat{v}_i \quad (2)$$

Each of the seven estimated A.C.E. Survey DSE terms (2), Each of these seven terms are, in fact, the sum of a REE and a DEE, respectively. The second-phase portion (the term in  $\hat{u}$ ) is a REE because the ratio in (3) is a “re-weighted expansion” of the second-phase estimate. In (2), the term in  $\hat{v}$  is a DEE because the estimate comprises a product of the second phase and the third-phase sampling weights. In (4) the DEE version of the third-phase sampling weight is rewritten as the equivalent REE following Kim et al (2000).

$$\alpha_{hgi} = \underbrace{w_{hi}^x}_{(DEE)} \underbrace{x_{hgi} \frac{\sum_{k \in A} w_{hk} x_{hgk}}{\sum_{k \in A_2} w_{hk} x_{hgk}}}_{(REE)}, \quad w_{hi} = \frac{N_h}{n_h} \quad (3)$$

$$\tau_c = \frac{\sum_{i \in A_2} s_{ic}}{\sum_{i \in A_3} s_{ic}} = \frac{\sum_{i \in A_2} \alpha_{hgi} s_{ic} \alpha^{-1}}{\sum_{i \in A_3} \alpha_{hgi} s_{ic} \alpha^{-1}} \quad (4)$$

### A.C.E. Variance Estimation Methodology

The A.C.E. Survey employed a jackknife replicate variance estimator for the DSE (Kim, et al, 2000; Starsinic and Kim, 2001). The DSE variance estimation system was designed to provide a convenient, general-purpose methodology that produced consistent jackknife replication variance estimates.

Each of the seven estimated DSE terms is replicated

using

(5), (6), and (7) which are derived from (2), (3), and (4). The  $j$  in parentheses designates the  $j$ th jackknife replicate wherein the  $j$ th first-phase sample block cluster is “deleted” and the estimated term is calculated without it:

$$\hat{y}^{(j)} = \sum_{i \in A_2} \alpha_{hgi}^{(j)} \hat{u}_i + \sum_{c=2}^3 \sum_{i \in A_3} \alpha_{hgi}^{(j)} \tau_c s_{ci} \hat{v}_i \quad (5)$$

$$\alpha_{hgi}^{(j)} = w_{hi}^{(j)} x_{hgi} \frac{\sum_{k \in A} w_{hk}^{(j)} x_{hgk}}{1} \quad (6)$$

$$\tau_c^{(j)} = \frac{\sum_{i \in A_2} \alpha_{hgi}^{(j)} s_{ic} \alpha^{-1}}{\sum_{i \in A_3} \alpha_{hgi}^{(j)} s_{ic} \alpha^{-1}} \quad (7)$$

The first step in implementing this variance estimation methodology was to calculate the replicate weights. Let the replicate weights after the first stage of sampling be the standard jackknife replicate weights:

$$w_{hi}^{(j)} = \begin{cases} 0 & \text{if } i = j \\ \frac{n_h}{n_h - 1} w_{hi} & \text{if } i, j \in h \\ w_{hi} & \text{otherwise} \end{cases} \quad (8)$$

Then, plugging (8) into (6) we obtain the following replicate weights:

$$\alpha_{hgi}^{(j)} = \begin{cases} 0 & \text{if } i = j \\ \frac{r_{hg}}{r_{hg} - 1} \frac{n_{hg} - 1}{n_{hg}} \frac{n_h}{n_h - 1} \alpha_{hgi} & \text{if } i \neq j, h_i = h_j, \\ & hg_i = hg_j, j \in A_2 \\ \frac{n_{hg} - 1}{n_{hg}} \frac{n_h}{n_h - 1} \alpha_{hgi} & \text{if } i \neq j, h_i = h_j, \\ & hg_i = hg_j, j \notin A_2 \\ \frac{n_h}{n_h - 1} \alpha_{hgi} & \text{if } i \neq j, h_i = h_j, hg_i \neq hg_j \\ \alpha_{hgi} & \text{if } i \neq j, h_i \neq h_j, hg_i \neq hg_j \end{cases} \quad (9)$$

$$RF_i^{(j)} = \frac{\alpha_{hgi}^{(j)}}{\alpha_{hgi}}$$

Note that this is an unusual form of the jackknife. Normally, the jackknife has as many replicates as block clusters. Here, we have 11,303 bcs,  $i$ , remaining after the second-phase of sampling. However, we must use replicates,  $j$ , equal in number to the first-phase sample size, 29,136 bcs. The bcs sampled out in the second-phase obviously do not contribute to the variance due to the second- and third-phases, but they must be included to accurately account for the variability due to the first-phase of sampling. "Deleting" a bc that was sampled out changes the weights of the other bcs that were in the same first phase sampling stratum.

Although (5) operates at the bc level, the appropriate person-level records must be summed to the bc level. In addition, the estimates of each DSE component term are made separately for each poststratum,  $r$ . For convenience, three indicator variables,  $a_p$ ,  $b_p$ , and  $c_p$ , were created to translate the three components of (5) into components based on person-level characteristics. The summation in (10) is over all persons,  $p$ , in a particular poststratum,  $r$ , in bc  $i$ . Let a TES *person* be a person in a housing unit flagged during A.C.E. housing unit matching as a potential geocoding error. Let (i)  $a_p = 1$  if *not* a TES person (ii)  $b_p = 1$  if TES person *and* from TES certainty bc and (iii)  $c_p = 1$  if TES person *and* from TES non-certainty bc.

$$\hat{y}_r^{(j)} = \sum_{i \in A_2} RF_i^{(j)} \sum_{p \in i \in r} \eta_p^{(j)} \alpha_{hgi} (a_p + b_p + c_p \tau_2^{(j)}) \quad (10)$$

$$\eta_p^{(j)} = \begin{cases} y_p = 0, 1 & \text{if person } p \text{ has resolved status} \\ \bar{y}_k^{(j)} & \text{if person } p \text{ has unresolved status} \end{cases}$$

The next step of the implementation is to adjust the certain imputation probabilities to account for replication. For some persons, their match, residence, or correct enumeration status,  $y_p$ , remains unresolved even after follow-up operations. In these cases, a probability for each unresolved status is imputed using an imputation cell technique, with each unresolved case in an imputation cell getting the same imputed probability (Ikeda, 2000). To incorporate the replication weight, we recalculate the imputed person status separately for each replicate,  $j$ , in imputation cell,  $k$ , as in the following:

$$\bar{y}_k^{(j)} = \frac{\sum_{\text{resolved } p \in k} RF_i^{(j)*} \alpha_{hgi} (a_p + b_p + c_p \tau_2^{(j)}) y_p}{\sum_{\text{resolved } p \in k} RF_i^{(j)*} \alpha_{hgi} (a_p + b_p + c_p \tau_2^{(j)})} \quad (11)$$

where  $\alpha_{hgi}^* = \alpha_{hgi}$  without non-interview adjustment

To compute the variance estimate for stratum  $r$ :

$$\text{Var}(\hat{DSE}_r) = \sum_j \frac{n_h - 1}{n_h} \left( \hat{DSE}_r^{(j)} - \hat{DSE}_r \right)^2 \quad (12)$$

where:

$$\hat{DSE}_r^{(j)} = \left( C_r - \Pi_r \right) \left( \frac{C\hat{E}_r^{(j)}}{\hat{N}_{n,r}^{(j)}} \right) \left( \frac{\hat{N}_{n,r}^{(j)} + \hat{N}_{i,r}^{(j)}}{\hat{M}_{n,r}^{(j)} + \left( \frac{\hat{M}_{o,r}^{(j)}}{\hat{N}_{o,r}^{(j)}} \right) \hat{N}_{i,r}^{(j)}} \right)$$

And finally, the variance of the national adjusted population estimate is:

$$\text{Var}(\hat{DSE}_{US}) = \sum_r \sum_{r'} \text{Cov}(\hat{DSE}_r, \hat{DSE}_{r'}) \quad \text{where} \quad (13)$$

$$\text{Cov}(\hat{DSE}_r, \hat{DSE}_{r'}) = \sum_j (\hat{DSE}_r^{(j)} - \hat{DSE}_r) (\hat{DSE}_{r'}^{(j)} - \hat{DSE}_{r'})$$

### Sampling strata collapsing

In 30 of the 691(US) final sampling strata (a combination of

the first- and second-phase sampling strata and the state) the situation arose where within the final sampling stratum (fss) a single bc was selected in the second-phase sampling from the first-phase sample of more than one bc. The variability due to second-phase sampling in these fss could not be measured by the jackknife replication method because dropping the single bc for replication caused the replicate factor to become either zero or undefined. To remedy this problem the 30 “problem” fss were collapsed to yield 664<sup>1</sup> (US) collapsed final sampling strata (cfss). The collapsing was done with bcs of similar sampling weights to lessen the sample weight variation and the associated variance increase in the cfss. Furthermore, each collapsing was done across states but involved only bcs in the same first phase stratum and second phase stratum (Starsinic, 2000b).

### Poststrata collapsing

The original (pre-collapsed) poststratification design (Haines, 2001) employed 64 poststrata groups each containing seven age/sex categories yielding 448 (US) poststrata, r. Eight of the 64 Poststrata groups had to be collapsed in order to reduce the variance of the DSEs for these poststrata. The collapsing was done within the poststrata groups by reducing the number of age/sex categories from seven to three. The determination of which poststrata groups to collapse was made based on two criteria. First, seven poststrata groups with small sample sizes (less than 100 persons) in at least one of the constituent age/sex categories were identified. Second, an eighth poststratum group was chosen for collapsing based on a post-data collection variance calculation which identified an outlier. The poststrata collapsing operation yielded 416 (US) collapsed poststrata, r, (56 poststrata groups x seven age/sex categories + 8 poststrata groups x three age/sex categories).

### Results

The variance estimates for nine comparison groups were calculated for both the 2000 A.C.E. (Table 2.) and the 1990 Post Enumeration Survey (PES) (Table 1.). These evaluation poststrata were chosen so that there would be exact correspondence between the persons defined for each group across the two surveys.

The comparison of cvs between the 1990 and the 2000 coverage surveys shows a decrease in relative standard

error of the total dual system estimate for all nine comparison groups. Furthermore, the results for Black and Hispanic Renters and American Indians on Reservations show a decrease in relative standard error of 40% or more.

**Table 1. 1990 PES CV(DSE) and DSE for 9 comparison groups**  
[n=165000 hus]  
[Excludes military and institution group quarters population]

Evaluation Poststratum	CV(DSE)	DSE
White/Other Race Owner	0.23%	134,696,700
White/Other Race Renter	0.49	53,358,400
Black Owner	0.57	13,730,500
<b>Black Renter</b>	<b>0.85</b>	16,664,900
Hispanic Owner	0.69	9,586,200
<b>Hispanic Renter</b>	<b>1.26</b>	12,474,300
Asian and Pacific Islander Owner	1.43	4,055,000
Asian and Pacific Islander Renter	2.43	3,346,900
<b>American Indian on Reservation</b>	<b>6.09</b>	425,100
Total	0.20	248,338,000

**Table 2. 2000 A.C.E. CV(DSE) and DSE for 9 comparison groups**  
[n=300913 hus]  
[Excludes all group quarters population]

Evaluation Poststratum	CV(DSE)	DSE
White/Other Race Owner	0.14%	148,331,800
White/Other Race Renter	0.31	47,512,500
Black Owner	0.46	16,660,600
<b>Black Renter</b>	<b>0.50</b>	17,550,200
Hispanic Owner	0.45	17,005,800
<b>Hispanic Renter</b>	<b>0.58</b>	18,546,300
Asian and Pacific Islander Owner	0.87	6,380,600
Asian and Pacific Islander Renter	1.03	4,294,200
<b>American Indian on Reservation</b>	<b>1.26</b>	567,100
Total	0.20	276,849,100

<sup>1</sup> Three of 30 “problem” fss were collapsed with the remaining 27 “problem” fss to yield 664 cfss =691 fss - 27 fss

**Table 3. 2000 A.C.E. and 1990 PES Sample Size**

Evaluation Poststratum	1990	2000
White/Other Race Owner	185,717	330,819
White/Other Race Renter	67,693	118,743
Black Owner	36,445	42,118
<b>Black Renter</b>	<b>34,940</b>	<b>46,947</b>
Hispanic Owner	19,504	42,406
<b>Hispanic Renter</b>	<b>21,177</b>	<b>53,239</b>
Asian and Pacific Islander Owner	4,698	17,162
Asian and Pacific Islander Renter	2,944	12,319
<b>American Indian on Reservation</b>	<b>3,887</b>	<b>13,665</b>
Total	377,005	677,418

### Discussion

The decrease in the CVs can be generally explained by the increase in sample size of approximately 82% from the 1990 to 2000. The three comparison groups : Black Renter, Hispanic Renter, and American Indian on Reservation, however, had a decrease in their CV(DSE) which appears to be beyond what can be explained by a larger sample size alone. The multi-phase A.C.E. design employed a second phase sampling plan which used the changes in the measures of size of sample bcs from the first phase to the second phase as a basis for stratification. Specifically, certain bcs that actually “jumped” size categories were selected with certainty in the second phase sampling. This decreased weight variation and thereby variance, especially in the three comparison groups cited above.

In conclusion, the variance estimation methodology used in the A.C.E. was designed to accurately measure the error in the DSE due to sampling, imputation of missing status, and TES sampling. This methodology allowed us to more accurately characterize the benefits of the complex A.C.E. design than would have been possible with, for example, a simple stratified jackknife.

### References

Cochran, W. G. (1977), *Sampling Techniques (3<sup>rd</sup> ed.)*, New York: Wiley.

Davis (2001), “Accuracy and Coverage Evaluation: Dual System Estimation Results”, Census 2000 Procedures and Operations Memorandum Series, B-9\*, February 28, 2001.

Haines, D. (2001), “Accuracy and Coverage Evaluation: Computer Specifications for Person Dual System

Estimation (U.S.)”, Census 2000 Procedures and Operations Memorandum Series, Q-48, March 12, 2001.

Ikeda, M. (2000), “Accuracy and Coverage Evaluation: Specifications for Missing Data Procedures”, Census 2000 Procedures and Operations Memorandum Series, Q-25, April 20, 2000.

Navarro, A., and Olson, D. (2001), “Accuracy and Coverage Evaluation: Effect of Targeted Extended Search”, Census 2000 Procedures and Operations Memorandum Series, B-18\*, February 28, 2001.

Kim, J.K., Navarro, A., and Fuller, W. (2000), “Replication Variance Estimation for Multi-Phase Stratified Sampling”, Internal Census Bureau memorandum, November 16, 2000.

Kott, P. S. and Stukel, D. M. (1997), “Can the Jackknife Be Used With a Two-Phase Sample”, *Survey Methodology*, **23**, 81-89.

Rao, J. N. K. (1973), “On double sampling for stratification and analytical surveys”, *Biometrika*, **60**, 125-133.

Rao, J. N. K. and Shao, J. (1992), “Jackknife variance estimation with survey data under hot deck imputation”, *Biometrika*, **79**, 811-822.

Starsinic, M. (2000a), “Accuracy and Coverage Evaluation: Overview of Census 2000 A.C.E. Variance Estimation - Theory and Implementation”, Census 2000 Procedures and Operations Memorandum Series, V-2, September 19, 2000.

Starsinic, M. (2000b), “Collapsing of Sampling Strata for A.C.E. Variance Estimation, Census 2000 Procedures and Operations Memorandum Series, V-, June 8, 2000.

Starsinic, M. and Kim, J.K. (2001), “Accuracy and Coverage Evaluation: Computer Specifications for Variance Estimation for Census 2000”, Census 2000 Procedures and Operations Memorandum Series, V-5, March 9, 2001.

Thompson, J. (1992), “CAPE Processing Results”, Internal Census Bureau memorandum, March 20, 1992.

Wolter, K. (1986), “Some Coverage Error Models for Census Data”, *Journal of the American Statistical Association*, **81**, 338-346.