Paper ST03

Variance Estimates for Census 2000 Using SAS/IML® Software

Peter P. Davis, U.S. Census Bureau, Washington, DC[1]

ABSTRACT

Large variance-covariance matrices are not uncommon in statistical data analysis. For Census 2000, we produced coverage estimates as part of the Accuracy and Coverage Evaluation (A.C.E.) Survey for 448 population subgroups called post-strata. Consequently, the A.C.E. variance estimation operation produced a 448 by 448 variance-covariance matrix for these national post-strata. In obtaining variance estimates of these population subgroups for analysis of the census, publication, and for future research, basic ideas of matrix theory and manipulation were essential in calculating accurate measures of reliability. SAS/IML® software allowed for quick and easy matrix multiplication and matrix operations to acquire the desired variance estimates for any combination of the 448 post-strata.

INTRODUCTION

For the Census 2000 A.C.E., we divided the population into population subgroups called post-strata. Post-stratification groups together people who have similar coverage within the census.

The 2000 A.C.E. post-stratification was based on seven variables: race, Hispanic origin, tenure, region, Metropolitan Statistical Area size, Type of Enumeration Area, and tract level return rate. These seven variables define the 64 major post-stratum groups as seen in Table 1. Within each post-stratum group, there are seven age/sex groups as seen in Table 2. Therefore, the post-stratification design for Census 2000 A.C.E. contained 448 post-strata resulting from the cross-classification of the 64 major post-stratum groups and the seven age/sex groups. This 448 national post-stratification plan was chosen to reduce correlation bias without having an adverse effect on the variance of the Dual System Estimator. (See Griffin and Haines, 2000.)

The Census 2000 A.C.E. Survey employed a dual-system model to estimate the true population. The dual system estimate (DSE) is a complex ratio estimator with multiple components. We calculated the DSE for all of the 448 post-strata. The variance estimation operation produced a 448 by 448 variance-covariance matrix of the DSEs. From this matrix, we could obtain measures of reliability for all of the national post-strata. These DSE estimates and their variances were published in Davis (2001).

In addition to the detailed Dual System Estimation computations, useful "roll-ups" that aggregate the DSE results by age and sex, tenure, minority/nonminority, or other summations were necessary to determine Census 2000 coverage. With such a large variance-covariance matrix, finding variance estimates for these "roll-ups" could be cumbersome. However, by using SAS/IML® and with a working knowledge of matrix theory and matrix manipulation, these "roll-ups" were calculated so that Census 2000 A.C.E. coverage estimates could be determined and also potentially compared with 1990 Census results.

MATRIX THEORY

Consider a random vector consisting of the 448 DSE observations V_1 through V_{448}. The variances of these random variables and the covariances between any two observations form the variance-covariance matrix *V*. The variance-covariance matrix *V* is a 448 by 448 matrix taking the following form:

$$V = \begin{bmatrix} v_{1,1} & \cdots & v_{1,448} \\ \vdots & \ddots & \vdots \\ v_{448,1} & \cdots & v_{448,448} \end{bmatrix} \qquad \textbf{(1)}$$

The entries on the main diagonal, $v_{i,i}$, are the variances and the off-diagonal entries are the covariances, $v_{i,j}$, with i … j. Remember, of course, that $v_{i,j} = v_{j,i}$ for all i … j. Hence, *V* is a symmetric matrix.

Consider the column vector **x**.

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{448} \end{bmatrix} \qquad \textbf{(2)}$$

where $x_1, x_2, \ldots, x_{448}$ are real numbers. We form a row vector by taking the transpose of a column vector. This is denoted by **x**′.

Recall the quadratic form c = **x**′*V***x**. At this point, the most important element of a quadratic form is its dimensions. Since **x** is a 448 by 1 column vector, **x**′ is a 1 by 448 row vector. Hence, c = **x**′*V***x** is a 1 by 1 matrix. Thus when $x_1, x_2, \ldots, x_{448}$ assume numeric values and *V* is the variance-covariance matrix of the DSE, also containing real numbers, then c is just a constant. This can be more readily seen when the quadratic form c is expanded as a sum of squares and cross products. The expanded form is:

$$c = \sum_{i=1}^{448} \sum_{j=1}^{448} x_i v_{ij} x_j \qquad \textbf{(3)}$$

Finally, some basic results in probability and statistics from Neter et al. (1996, p. 1,318.)

---

[1] The author is a mathematical statistician in the Decennial Statistical Studies Division, U.S. Census Bureau. 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.

Let $V_1, ..., V_n$ be n random variables. Consider the function $\sum a_i V_i$, where the a_i are constants. We then have:

$$\sigma^2\left\{\sum_{i=1}^{n} a_i V_i\right\} = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \sigma\{V_i, V_j\} \qquad \textbf{(4)}$$

For n = 2, we have:

$$\sigma^2\{a_1V_1 + a_2V_2\} = a_1^2\sigma^2\{V_1\} + a_2^2\sigma^2\{V_2\} + 2a_1a_2\sigma\{V_1, V_2\} \qquad \textbf{(5)}$$

THE MATRIX STRUCTURE

Before calculating any summary DSE variance results, for example, for the seven age/sex groups or for tenure (owner and renter), it is necessary to understand the structure of the DSE variance-covariance matrix. The layout of the DSE variance-covariance matrix combines the post-stratum groups from Table 1 with the age/sex groups of Table 2. As mentioned before, there are 64 major post-stratum groups for the Census 2000 A.C.E. Within each of the 64 post-stratum groups there are seven age/sex groups comprising 64 × 7 = 448 post-strata.

So, for example, $v_{1,1}$ is the variance of post-stratum group 1 who are under 18 years old. $v_{2,2}$ is the variance of post-stratum group 2 who are males 18 to 29 years of age. Similarly, $v_{448,448}$ is the variance of post-stratum group 64 who are females 50 years of age and older.

Looking at Table 1, what if we desired the DSE variance estimate for Domain 7: Non-Hispanic White or "Some other race." To obtain this variance estimate, we require the sum of all the elements within the first 280 rows and the first 280 columns of the DSE variance-covariance matrix *V*. In essence, we are creating a submatrix of *V* containing only variances and covariances relevant to Domain 7.

METHODOLOGY

SAS/IML® has a function to subset a matrix and then sum all of the elements within that submatrix. However, consider a column vector **x** of 0s and 1s. Let **x** be 448 by 1 such that the first 280 elements within the column vector **x** are a 1 and the remaining 168 are 0s.

$$x = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

Then, from the matrix theory described above, the quadratic form $c = \mathbf{x}'V\mathbf{x}$ would equal the variance of the DSE for Domain 7: Non-Hispanic White or "Some other race." More thoroughly, if we expand the quadratic form using equation (3), we get:

$$c = \sum_{i=1}^{280} 1^2 v_{i,i}^2 + 2\sum_{i=1}^{280}\sum_{\substack{j=1 \\ (j>i)}}^{280} 1^2 v_{i,j} \qquad \textbf{(6)}$$

It can be shown, that by expanding equation (4) with n = 448, a_1 through a_{280} = 1, and a_{281} through a_{448} = 0, we achieve the result in equation (6).

Variance estimates for each of the remaining six race/origin domains can be obtained in a similar fashion. For any summary DSE variance result, we are, in essence, partitioning the DSE variance-covariance matrix V by using a column vector **x** of 0s and 1s. For each summary result desired, only the design of column vector **x** changes. Placement of the 0s and 1s within the vector **x** is critical to acquiring the desired summary variance result.

CALCULATING VARIANCES USING SAS/IML®

The following sets of SAS code outline how to manipulate the variance-covariance matrix to obtain several summary DSE variance results.

Domain 7: Non-Hispanic White or "Some other race"

Beginning with Domain 7: Non-Hispanic White or "Some other race," we have the following:

```
proc iml;
/* Read in Var-Cov Matrix.   */
use vdse;
/* Var-Cov Matrix stored in VCDSE.  */
read all var _num_ into vcdse;
/* Create default column vector of 0s. */
x=j(448,1,0);
/* Do-Loop to assign x = 1 for all i in
Domain 7*/
do i = 1 to 280;
  x[i,1]=1;
end;
domain7=x`*vcdse*x;
print domain7;
quit;
```

VDSE is the SAS data set containing the DSE variance-covariance matrix. The matrix is read into VCDSE. VCDSE is 448 by 448. The "J Function" in SAS/IML® creates a matrix of identical values. Here we create a 448 by 1 column vector **x** of 0s. In the Do-Loop, we assign a 1 to the first 280 elements of **x**. Now we have our desired column vector **x** corresponding to the rows and columns of Domain 7 in the DSE variance-covariance matrix. Then DOMAIN7 is the quadratic form described above which gives the variance of the DSE for Domain 7: Non-Hispanic White or "Some other race." Similar code can be written to partition the variance-covariance matrix for any of the remaining six race/origin domains.

The Nation
With a slight adjustment, we can find the variance of the DSE for the entire nation. This is accomplished by creating a column vector **x** of 1s.

```
proc iml;
use vdse;
read all var _num_ into vcdse;
/*  Create National column vector.   */
x=j(448,1,1);
natl=xN*vcdse*x;
print natl;
quit;
```

Then NATL is the quadratic form which gives the variance of the DSE for the entire nation.

Tenure: Owner and Renter
Summary results for the DSE get a little more complicated when the desired rows and columns are no longer adjacent. For example, suppose we were interested in the variance of the two Tenure Groups: Owner and Renter. Notice in Table 1, Owner and Renter are spread out among the 64 major post-stratum groups. Defining the appropriate rows and columns requires a little more care.

For instance, let's focus on Owner. Owner falls in post-stratum groups 1 through 32, 41 through 44, 49 through 52, 57, 59, 61, and 63. Now the appropriate rows and columns must be identified. Keeping in mind each post-stratum group contains seven age/sex categories, then post-stratum 1 through post-stratum 32 correspond to rows and columns 1 through 224. The remaining rows and columns pertaining to Owner are found in a similar manner. The code for calculating the variance Owner follows:

```
proc iml;
use vdse;
read all var _num_ into vcdse;
/* Create default column vector of 0s. */
x=j(448,1,0);
/* Do-Loops to assign x = 1 for all i in
Owner */
do i = 1 to 224;
  x[i,1]=1;
end;
do i = 281 to 308;
  x[i,1]=1;
end;
do i = 337 to 364;
  x[i,1]=1;
end;
do i = 393 to 399;
  x[i,1]=1;
end;
do i = 407 to 413;
  x[i,1]=1;
end;
do i = 421 to 427;
  x[i,1]=1;
end;
do i = 435 to 441;
  x[i,1]=1;
end;
owner=xN*vcdse*x;
print owner;
quit;
```

Then OWNER is the quadratic form which gives the variance of the DSE for owners. The variance for renters is found using the same technique.

Age/Sex Groups
Obtaining the variances for the seven age/sex groups is a little more tedious. Within the variance-covariance matrix, same age/sex groups are seven rows apart; that is, in order to calculate the variance for children under the age of 18, you need the information from row 1 and column 1, row 8 and column 8, row 15 and column 15, etc.... Starting with the first element, the column vector **x** would have a one in every seventh position and then zeros elsewhere. Similarly, for males age 18 to 29, starting with the second element, the column vector **x** would have a one in every seventh position and then zeros elsewhere. Continuing in this same pattern and by using the matrix multiplication of SAS/IML®, we obtain the variances for the seven age/sex groups.

RESULTS
The results of this work were presented to the Executive Steering Committee on A.C.E. Policy at the U.S. Census Bureau to assist them in assessing the Census 2000 data with and without statistical adjustment. The results are outlined in Davis (2001). Table 3 displays some of these results.

DISCUSSION
One could argue, that while most programming languages deal with single data elements, since the fundamental data element in SAS/IML® is the matrix, my task became much easier. The built-in matrix operations in SAS/IML® were essential to the work I was doing which needed completion in a shortened time frame. Fortunately, I was made aware of them at the right time.

REFERENCES

Davis, P. (2001), "Accuracy and Coverage Evaluation: Dual System Estimation Results," DSSD Census 2000 Procedures and Operations Memorandum Series B-9*, U.S. Census Bureau, Washington, DC.

Griffin, R. and Haines, D. (2000), "Accuracy and Coverage Evaluation Survey: Final Post-stratification Plan for Dual System Estimation," DSSD Census 2000 Procedures and Operations Memorandum Series Q-24, U.S. Census Bureau, Washington, DC.

Neter, J. et al. (1996). Applied Linear Statistical Models, 4th ed., IRWIN, Chicago, Illinois.

Ortega, J. (1991). *Matrix Theory: A Second Course*, 3rd ed., Plenum Press, New York, New York.

SAS Institute Inc. (1999). *SAS/IML® User's Guide*, Version 8, SAS Publishing, Cary, NC.

Starsinic, M. (2001), “Accuracy and Coverage Evaluation Survey: Specifications for Covariance Matrix Output Files from Variance Estimation for Census 2000,” DSSD Census 2000 Procedures and Operations Memorandum Series V-4, U.S. Census Bureau, Washington, DC.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Peter Davis
U.S. Census Bureau
4700 Silver Hill Road, Stop 7600
Washington, DC 20233-7600
Work Phone: 301.457.8322
Fax: 301.457.4580
Email: peter.p.davis@census.gov
Web: www.census.gov

Table 1: Census 2000 A.C.E. - 64 Post-Stratum Groups

Race/Origin Domain Number*		Tenure	MSA/TEA	High Return Rate				Low Return Rate			
				N	M	S	W	N	M	S	W
Domain 7 (Non-Hispanic White or "Some other race")		Owner	Large MSA MO/MB	01	02	03	04	05	06	07	08
			Medium MSA MO/MB	09	10	11	12	13	14	15	16
			Small MSA & Non-MSA MO/MB	17	18	19	20	21	22	23	24
			All Other TEAs	25	26	27	28	29	30	31	32
		Non-Owner	Large MSA MO/MB	33				34			
			Medium MSA MO/MB	35				36			
			Small MSA & Non-MSA MO/MB	37				38			
			All Other TEAs	39				40			
Domain 4 (Non-Hispanic Black)		Owner	Large MSA MO/MB	41				42			
			Medium MSA MO/MB								
			Small MSA & Non-MSA MO/MB	43				44			
			All Other TEAs								
		Non-Owner	Large MSA MO/MB	45				46			
			Medium MSA MO/MB								
			Small MSA & Non-MSA MO/MB	47				48			
			All Other TEAs								
Domain 3 (Hispanic)		Owner	Large MSA MO/MB	49				50			
			Medium MSA MO/MB								
			Small MSA & Non-MSA MO/MB	51				52			
			All Other TEAs								
		Non-Owner	Large MSA MO/MB	53				54			
			Medium MSA MO/MB								
			Small MSA & Non-MSA MO/MB	55				56			
			All Other TEAs								
Domain 5 (Native Hawaiian or Pacific Islander)		Owner		57							
		Non-Owner		58							
Domain 6 (Non-Hispanic Asian)		Owner		59							
		Non-Owner		60							
American Indian or Alaska Native	**Domain 1** (On Reservation)	Owner		61							
		Non-Owner		62							
	Domain 2 (Off Reservation)	Owner		63							
		Non-Owner		64							

*For Census 2000, persons can self-identify with more than one race group. For post-stratification, persons are included in a single Race/Origin domain. This does not change a person's actual response and all persons weree tabulated based on their actual response in the census.

Table 2: Census 2000 A.C.E. - 7 Age/Sex Groups

Age	Male	Female
Under 18	1	
18 to 29	2	3
30 to 49	4	5
50+	6	7

Table 3: Census 2000 A.C.E. DSE Summary Results for Major Groups

Census 2000 A.C.E.*	
Characteristic	**Standard Error of the DSE**
Total	377,918.25
Race/Origin Domain	
Non-Hispanic White	272,254.20
AI Off Reservation	22,295.65
Non-Hispanic Black	120,934.82
Hispanic	140,523.57
Non-Hispanic Asian	64,933.64
Hawaiian or Pacific Isl.	17,959.10
AI On Reservation	7,132.18
Tenure	
Owner	263,625.19
Renter	234,514.48
Age/Sex	
0-17	140,765.41
18-29 Male	75,607.51
18-29 Female	65,466.22
30-49 Male	82,611.61
30-49 Female	73,120.22
50+ Male	61,050.66
50+ Female	66,666.48

*The Census 2000 A.C.E. Dual System Estimate standard errors in this report are for the household population.