This paper describes an estimation and sampling errors package for use with survey data consisting of two SAS procedures (NASSTIM and NASSVAR) and a SAS preprocessor. Survey estimates and their associated sampling error statistics are computed for user-specified characteristics using balanced repeated half-sample replications (BRR) of the full sample.

The SAS preprocessor prepares the input survey data file for use by PROCES NASSTIM and NASSVAR. The preprocessor half sample replicates to which an observation belongs and optionally incorporates user-specified ratio adjustments into "weight" factors which are permanently affixed to each survey observation.

The procedures NASSTIM and NASSVAR perform the actual estimation and sampling error computations upon the survey data file. Both PROCES allow the user to specify arithmetic expressions to create "computed estimates" for which the full complement of sampling error statistics are produced.

PROC NASSTIM computes estimates and missing value statistics for specified characteristics while NASSVAR computes sampling errors and associated statistics in addition to the estimate.

A knowledgeable statistician can use this package as an effective tool. The user must be familiar with the sample design and the methods of computing sampling errors; specifically, the user must determine: (a) the appropriate methods of defining half samples that, except for the sample size, simulate the original design and properly reflect all stages of sampling and estimation; (b) the number of half-sample replications desired for variance estimation and prepare the codes that define them; (c) the modifications necessary in the full sample estimation procedure that may be called for by samples half as large; and (d) the method for defining the records and weights to be used for each of the half samples. The use of the package is illustrated by an example.

The software discussed in this paper can be adapted by the user to reflect weighting systems for full sample estimates, and for replicated half samples as well; thus, sampling error estimates for complex weighting systems may be computed. Any stratified sample design featuring a selection of pairs of primary sampling units (PSU's) from each stratum may be accommodated; self-representing PSU's can be adapted by designating pairs of half samples within each of these PSU's.

**DESIGN OBJECTIVES**

Requirements in developing the estimation and variance software were: (a) to provide estimates and sampling errors for statistics involving unspecified transformations (for example, sums, ratios, differences, logarithms of ratios) of the survey variables; (b) the incorporation of user-specified ratio adjustments into replicate level "weight" factors; (c) ease of use by nontechnical personnel; and (d) low per use cost.

**ESTIMATION OF SAMPLING ERRORS**

The package uses the Balanced Half-Sample Repeated Replication (BRR) method of variance estimation. This method was chosen for its generality and its ease of use. Variance can be estimated for a wide variety of statistics of interest, linear or nonlinear. This paper does not discuss the BRR method in detail; the subject is covered in a number of publications. (See McCarthy [1] and [2].) The programs supplied for the illustration have been written to incorporate several practical problems faced in analyzing the results of surveys using the BRR method.

The application of the BRR method for a design having two PSU's selected from each stratum involves the repeated re-estimation of the statistics using one half of the full survey PSU's. Each half sample contains one of the two PSU's selected from each stratum. If there are R half samples, then R + 1 estimates are prepared; one for each of the R half samples and the estimate based on the full sample.

**PREPARING ESTIMATES FROM SAMPLE DATA**

The estimation module, NASSTIM, is used to prepare weighted estimates from the sample records. NASSVAR, the sampling error module, measures the precision of the complete estimation procedure by applying the estimator to each of the replicated half samples. To approximate the effect of the precision of such adjustments, the factors in the ratio estimate should be recomputed separately for each half sample.

Although the computation of separate ratio adjustment factors for each half sample appears to be formidable, the factors can be obtained easily through a series of matrix operations. Appendix A contains a development of the computations using operators available in the SAS procedure MATRIX. The matrices which are required include a design matrix defining the PSU's belonging to each half sample and a series of totals used as the numerators and denominators of the ratio factors. The definitions of these matrices and several others needed for the computations are given in Appendix A.

The procedure of precomputing and attaching weights permanently has several advantages over the approach of computing the factors for each computer request. First, a reduced cost of processing can be expected since the constant recomputation of weights is avoided. Second, the user has a file of the survey results with half-sample weights attached.
EXECUTING PROC NASSTIM AND PROC NASSVAR

This section describes the SAS procedures, NASSTIM and NASSVAR. The procedure grammar is similar to that used in standard SAS procedures. The user must specify the variables for which estimates are to be computed, the weight variables to be used and optionally, any transformations required. The full range of SAS arithmetic operators and functions may be used in specifying the computation of new variables from estimates computed by the procedures. Estimates and their associated statistics may be computed for any number of subgroups of the input file through the use of a "BY" statement.

The procedures NASSTIM and NASSVAR both compute estimates of user-specified characteristics. The procedures prepare estimates as weighted totals; however, the user specifies the weights to be employed when the procedure is invoked. The procedures therefore can accommodate the selection of specific characteristics for estimation, and, on separate executions of the procedure, the use of various estimation methods. NASSTIM and NASSVAR prepare estimates and NASSVAR, in addition, computes sampling errors and variances for each given characteristic.

NASSVAR prepares estimates for each characteristic specified by the user within each of the half-sample replicates required for the sample design and also for the full sample. (The full sample is called replication zero.) After these estimates are calculated, the following statistics are displayed for each characteristic:

1. Estimate;
2. Number of cases missing for replication zero;
3. Weighted number of missing cases for replication zero;
4. Relvariance (ratio of variance to the square of estimate);
5. Variance;
6. Standard error;
7. Approximate lower 95 percent confidence bound;
8. Approximate upper 95 percent confidence bound; and
9. Coefficient of variation (%) (square root of relvariance).

The half-sample replicate to which a particular observation belongs is defined implicitly to NASSVAR through the use of a WEIGHT statement. The presence and order of variables (representing weights) on the WEIGHT statement define for NASSVAR the half-sample replicates to which a particular observation belongs. These weights are precomputed to reflect the estimation procedure employed. The user identifies the weights to be used and, by this process, defines the estimation procedure.

NASSVAR by default produces sampling errors based upon estimates of TOTAL variance; to produce estimates of sampling errors within first-stage sampling units, the user must specify the "WITHIN" procedure option and supply the set of weights which define half samples within each first-stage unit. In the example, the weights associate odd-numbered cases separately from even-numbered cases.

Both procedures have been designed to produce total estimates, ratios of estimates and almost any other arithmetic function available to SAS computed from the estimates. NASSVAR also computes associated sampling errors for these computed estimates. These arithmetic evaluations may include any estimates specified on the COMPVAR statement and/or user specified constants. The results of these computations are stored in variables specified on the OUTPUT statement discussed below. (User-specified constants may be included in expressions to allow a third stage of ratio adjustments.)

The annotated listing in Figure 1 provides an example of running the SAS procedures NASSTIM and NASSVAR. Three national estimates are computed from the NASS 1979 Analysis File: total number of urban accidents, total number of accidents and the ratio of the first to the second of these two estimates.

To obtain an estimate of the total number of accidents, a dummy variable, ACES, is set equal to "1" for all accident records (see line 3).

To obtain an estimate of the total number of urban accidents, a dummy variable, URBAN, is defined in lines 4-5. This variable is a recording of the NASS variable A21. URBAN is set equal to "1" for all accident records describing urban accidents and zero otherwise. NASSTIM and NASSVAR sum the weights of each of these created "dummy variables" (NASSVAR sums the weighted estimates at the replicate level).

To obtain an estimate of the proportion of URBAN accidents to total accidents, a new variable is created by the procedures; U_RATIO is defined (see lines 14 and 21) as the ratio of URBAN to ACES.

The output resulting from invoking NASSTIM (lines 9-15) is shown in Figure 2. Three lines are displayed; one for each statistic. The sum of the appropriate weights is given for ACES and for URBAN and their ratio is given as U_RATIO. Also shown are the number of records coded as missing for each variable and the weighted sum of the missing records.

The call to PROC NASSVAR appears in lines 16-22 and is similar to PROC NASSTIM. In the example, NASSVAR is to compute estimates and sampling errors for the same characteristics as NASSTIM. The output resulting from invoking NASSVAR is shown in Figure 3.

An output file, TWO, is constructed which will contain one record for each replicate; the record contains an estimate for each user-specified statistic based on one half sample.
Line 18 provides the names of the weight variables required. Since this example applies to a survey using 10 PSU's, five paired strata are necessary with eight replicates to be generated. Thus, nine weights are required; the first weight "R WGTD", is used to estimate replicate zero, the full sample estimate. Line 21 defines a transformation, the ratio of urban accidents to total accidents. At the user's option (lines 23-24), the output data set can be printed (Figure 4). Figure 4 contains nine lines, one for each replicate (REPL_ID). By using this option, the full sample estimate and the half-sample results can be examined for conformity. Unusual values, outliers, might suggest further investigation in certain PSU's present in an unusual half sample.

APPENDIX A

COMPUTATIONS REQUIRED FOR FIRST- AND SECOND-STAGE RATIO ADJUSTMENTS

This appendix describes in matrix terms the computations required to obtain ratio adjustment factors to reduce within- and between-PSU variances. Factors are developed for the full sample (replicate zero) and for two sets of half samples: one for total variance and one for within variance (replicates one through "k11). The numbers of half-sample replicates vary by sample design.

A.1 We begin by defining the following matrices used in the computation:

\[ A_{2hxu} \]

Known stratum totals for the variables selected to adjust each record type to reduce between-PSU variance.

\[ F_{2hxu} \]

Estimated totals for the variables contained in the A matrix (e.g., estimates of A obtained by multiplying PSU Census totals by the PSU weight).

\[ C_{2hxu} \]

Census counts of total records appearing in the sampled PSU weighted up to estimate the stratum total. These are used in the numerator of the ratio factors prepared to reduce the within-variance component.

\[ D_{2hxu} \]

Estimated stratum total numbers of odd and even records for the categories contained in the C matrix.

\[ F_{2hxc} \]

Matrix of 1's and 0's defining groupings of PSU's used to reduce between-PSU variances. Columns of F add to a vector of 1's.

\[ C_{2hkg} \]

Matrix of 1's and 0's defining groupings of PSU's used for reducing within-PSU variances. Columns of G add to a vector of 1's.

\[ R_{2hx(k+1)} \]

Matrix of 0's, 1's, and 2's defining replicates. First column is all 1's (replicate zero) and remaining columns are 0's or 2's defined for variance computation.

\[ I_F \]

Matrix of 1's and 0's used to collapse over c between-PSU ratio groups:

\[
\begin{bmatrix}
1 & 0 & 0 & \ldots & 0 & 0 & 0 & \ldots & 0 \\
0 & 1 & 0 & \ldots & 0 & 0 & 0 & \ldots & 0 \\
& & & \ddots & & & & \ddots & & \\
0 & 0 & 0 & \ldots & 1 & 0 & 0 & \ldots & 0 & 0
\end{bmatrix}
\]

\[ I_C \]

Matrix of 1's and 0's used to collapse over g between-PSU ratio groups (see 1p).

\[ I_{7hx(k+1)} \]

Matrix of 1's used to reproduce F in step 1b in A.4 below.

A.2 The subscripts used in A.1 are defined below:

- **h** = number of strata; 2h indicates two rows per PSU, first for odd cases, second for even cases within PSU. Note that for estimates of total variance, the cases are combined in nonself-representing PSU's.

- **u** = number of record types used in preparing within-PSU ratio factors.

- **c** = number of groups of PSU's for which between-PSU factors are prepared.

- **g** = number of groups of PSU's for which within-PSU factors are prepared.

- **k** = number of half-sample replicates required. Different for estimating total variance and within-variance for the various sample designs.

A.3 Several types of matrix operations are defined below.

- **a. Matrix term-by-term addition:**

  \[ M_{axb} + N_{axb} = O_{axb} \]

- **b. Matrix term-by-term multiplication:**

  \[ M_{axb} * N_{axb} = O_{axb} \]

- **c. Matrix transpose:**

  \[ M'_{hxa} = M_{axh} \]

- **d. Matrix multiplication (dot products):**

  \[ M_{axb} * N_{bxc} = O_{axc} \]
e. Horizontal direct product:

\[ M \times N = O \]

**NOTE:** In SAS, the result of this operation is a matrix with the same number of rows as \( M \) and \( N \), and a number of columns equal to the product of the number of columns of \( M \) times the number of columns of \( N \).

f. Matrix term-by-term division:

\[ M \div N = 0 \]

### A.4 Using the above definitions, the required computations are given below.

1. **Compute factors to reduce the between-PSU variance:**

   a. **Compute between-PSU design matrix:**

   \[ J_{2h}(x(k+1)) = R_{2h}(x(k+1)) \]

   b. **Compute between-PSU numerator terms:**

   \[ \text{NUM}_{x(k+1)} = A'_{x2h} \times (2h_{x(k+1)} \otimes F_{2hx}) \]

   c. **Compute between-PSU denominator terms:**

   \[ \text{DEN}_{x(k+1)} = B'_{x2h} \times J_{x(k+1)} \]

   d. **Compute between-PSU ratio terms:**

   \[ \text{RATIO}_{x(k+1)} = \text{NUM}_{x(k+1)} \div \text{DEN}_{x(k+1)} \]

   e. **Expand between-PSU ratio factors over all PSU by record type cells in the between-PSU design matrix:**

   \[ \text{PROD}_{(k+1)x2h} = \text{RATIO}_{x(k+1)} \]

   \[ \otimes | \text{SORT}(J')_{x(k+1)x2h} \]

   **NOTE:** The square root is taken since two multiplicative adjustment factors will be computed (one for between and one for within) and each contains a factor of two used with each half sample.

2. **Compute factors to reduce the within-PSU variance.**

   Repeat 1a through 1f above to produce within-PSU factors by substituting as follows:

   - **G** for \( A \)
   - **D** for \( B \)
   - **G** for \( F \)
   - **I_G** for \( I_F \)
   - **F_W** for \( F_B \)

   In step 1b, substitute \( R \otimes I \) for \( I \otimes F \) because (contrary to the between-PSU factors) the numerators of within-PSU factors vary by replicate.

3. **Compute final ratios as:**

   \[ F_{(k+1)x2h} = F_W \otimes F_B \]

### BIBLIOGRAPHY


NOTE: SAS OPTIONS SPECIFIED ARE:
SORT=1

DATA ONE
1
2     SET INITIATORS 1
3     ACCS=1 1
4     IF A21=2 THEN URBAN=1 1
5     ELSE URBAN=0 1
6     LABEL
7     ACCS=ACCIDENTS
8     URBAN=URBAN ACCIDENTS
9
NOTE: DATA SET WORK.ONE HAS 3331 OBSERVATIONS AND 60 VARIABLES. 83 OBS/TRK.
NOTE: THE DATA STATEMENT USED 1.28 SECONDS AND 192K.

PROC NASSIM DATA=ONE 1
10    VAR ACCS URBAN 1
11    WEIGHT R_WGTO 1
12    COMPRVAR ACCS URBAN 1
13    OUTVAR U_RATIO 1
14    U_RATIO = URBAN / ACCS 1
15    TITLE SELECTED NATIONAL LEVEL ACCIDENT CHARACTERISTICS 1
16
NOTE: NASSIM IS AN UNSUPPORTED, EXPERIMENTAL PROCEDURE.
WESTAT INC
1050 RESEARCH BLVD
ROCKVILLE, MD 20850
(301) 251-1500

NOTE: THE PROCEDURE NASSIM USED 1.63 SECONDS AND 192K AND PRINTED PAGE 1.

PROC NASSVAR DATA=ONE TOTAL OUTPUT OUTDATA=TWO 1
16    VAR ACCS URBAN 1
17    WEIGHT R_WGTO-R_WGTO 1
18    COMPRVAR ACCS URBAN 1
19    OUTVAR U_RATIO 1
20    U_RATIO = URBAN / ACCS 1
21    TITLE SELECTED NATIONAL LEVEL ACCIDENT CHARACTERISTICS 1
22
NOTE: NASSVAR IS AN UNSUPPORTED, EXPERIMENTAL PROCEDURE.
NOTE: DATA SET WORK.TWO HAS 9 OBSERVATIONS AND 4 VARIABLES. 529 OBS/TRK.
WESTAT INC
1050 RESEARCH BLVD
ROCKVILLE, MD 20852
(301) 251-1500

NOTE: THE PROCEDURE NASSVAR USED 4.52 SECONDS AND 200K AND PRINTED PAGE 2.

PROC PRINT DATA=TWO 1

Figure 1.
### Figure 2.

<table>
<thead>
<tr>
<th>NAME</th>
<th>LABEL</th>
<th>ESTIMATE</th>
<th>MISSING</th>
<th>WEIGHTED MISSING</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACCS</td>
<td>URBAN URBAN ACCIDENTS</td>
<td>6704645</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>U_RATIO</td>
<td>COMPUTED ESTIMATE</td>
<td>4674048</td>
<td>0</td>
<td>N/A</td>
</tr>
</tbody>
</table>

### Figure 3.

<table>
<thead>
<tr>
<th>OBS</th>
<th>REPL_ID</th>
<th>ACCS</th>
<th>URBAN</th>
<th>U_RATIO</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>6704645</td>
<td>4674048</td>
<td>0.697136</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>6021238</td>
<td>3154635</td>
<td>0.523951</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>6605781</td>
<td>4514185</td>
<td>0.683369</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>7260034</td>
<td>5317400</td>
<td>0.731434</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>7387759</td>
<td>6314438</td>
<td>0.620426</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
<td>6572666</td>
<td>4754515</td>
<td>0.723399</td>
</tr>
<tr>
<td>7</td>
<td>6</td>
<td>6476077</td>
<td>3668766</td>
<td>0.566336</td>
</tr>
<tr>
<td>8</td>
<td>7</td>
<td>6650030</td>
<td>5263451</td>
<td>0.791493</td>
</tr>
</tbody>
</table>

### Figure 4.

<table>
<thead>
<tr>
<th>OBS</th>
<th>REPL_ID</th>
<th>ACCS</th>
<th>URBAN</th>
<th>U_RATIO</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>6704645</td>
<td>4674048</td>
<td>0.697136</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>6021238</td>
<td>3154635</td>
<td>0.523951</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>6605781</td>
<td>4514185</td>
<td>0.683369</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>7260034</td>
<td>5317400</td>
<td>0.731434</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>7387759</td>
<td>6314438</td>
<td>0.620426</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
<td>6572666</td>
<td>4754515</td>
<td>0.723399</td>
</tr>
<tr>
<td>7</td>
<td>6</td>
<td>6476077</td>
<td>3668766</td>
<td>0.566336</td>
</tr>
<tr>
<td>8</td>
<td>7</td>
<td>6650030</td>
<td>5263451</td>
<td>0.791493</td>
</tr>
</tbody>
</table>

### Figure 4.