AUTOMATION OF SAMPLING ERROR COMPUTATIONS FOR COMPLEX ESTIMATORS

David Morganstein, Robert H. Hanson, and Greg Binzer
Westat, Inc.

1. INTRODUCTION

This paper describes a computer package which operates in the Statistical Analysis System (SAS) to compute sampling errors using balanced repeated half-sample replications (BRR). An important requirement in developing the package was to provide estimates and sampling errors for statistics involving unspecified transformations (for example sums, ratios, differences, logarithms of ratios) of the survey variables. Another objective was to incorporate the use of ratio adjustments applied separately to two stages of sampling into the BRR computation of error estimates. A third feature of the approach was to permanently affix weight factors used for the estimation and error computations to the sampling unit records, thereby permitting users of the survey data to compute sampling errors for their estimates.

An interesting computational result presented deals with the preparation of a separate set of ratio-adjusted weights for each half-sample. Although the computations appear extensive, it is shown that these computations are succinctly expressed in matrix terms. By using a language or computer package containing matrix operations, these numerous computations can be accomplished with little effort.

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 can be accommodated; self-representing PSU's can be adapted by designating pairs of half samples within each of these PSU's.

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 simulate the original design and reflect all stages of sampling and estimation; (b) the number of half-sample replications desired for variance estimation; (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 system involves an estimation module, called NASSTIM, and a sampling errors module, called NASSVAR. An initial pre-processing operation is also needed to organize the data file as required for NASSTIM and NASSVAR. At a minimum, the estimation module requires the data file to show the selection probability for each record; the sampling errors module requires sufficient information to enable assignment of records to half samples. Records comprising the half samples are defined by a matrix of replication codes to specify records that make up the half samples; the matrix of codes is supplied by the user. The user must also supply and incorporate ancillary information used in estimation procedures (ratio estimate factors for the full sample and for each replicate, for example).

The NASSTIM estimation module prepares weighted totals of the records and transformations of the estimated totals. The transformations may be any of a wide range of the transformations contained in the SAS package. If desired, the estimation module may incorporate factors to modify the weights that appear on the records, as in ratio estimation. In the example given in this paper, we illustrate a weighting system involving two stages of ratio estimation. Additional stages of estimation may be added.

The NASSVAR sampling error computation module performs all of the computations of the estimation module and repeats the estimation computations for each half sample. The half-sample estimates are used to compute a variance about the full-sample estimate. The individual half-sample estimates are displayed for outlier examination and, as an option, are saved in a SAS data set.

2. AN ILLUSTRATION, THE NASS PROGRAM

The application of the package is explained by the example of the National Accident Sampling System (NASS) (2). This survey was designed to provide a representative source of accident data for use by regulatory agencies and accident researchers. The following is a brief summary of pertinent elements of the sample design and estimation procedure.

The universe from which cases are selected for investigation consists of motor vehicle accidents occurring on public highways for which a police accident report is completed and filed. The survey is not intended to represent other vehicle accidents. The sample is selected in several steps. First, a sample of one primary sampling unit (PSU) is selected from each of a set of strata into which the PSU's are grouped; PSU's are counties or groups of counties. Second, police jurisdictions in each sample PSU are stratified by size and an equal probability sample of jurisdictions is selected within the jurisdiction size strata. Finally, accidents reported at the sample jurisdictions are listed and a systematic sample of accidents is selected within the sample jurisdiction.

National estimates of accident data may be produced by an inflation estimator; that is, by weighting the sample records by the inverse of the probability of selection. Ratio adjustments are performed to reduce the variance introduced at the several steps in sampling.

The NASS estimation procedure routinely uses two stages of ratio estimation. At the first stage, data known about all PSU's in a stratum are used to reduce the between-PSU variance.

Similarly, variables known about all police jurisdictions are used to reduce the within-PSU sampling errors. Further stages of ratio estimation can also be introduced using characteristics obtained from independent sources for which consistent estimates can be prepared from the NASS that are expected to be correlated with the key variables of interest.

3. PREPARING ESTIMATES FROM SAMPLE DATA

The estimation module, NASSTIM, is used to prepare weighted estimates from the sample records. 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. This recomputation can be tedious and expensive if special purpose programs must be written to perform the work. A separate program, part of the pre-processing operation, performs these computations; in the version provided, the user must supply only the known subgroup totals.

The programs compute adjustment factors as a pre-processing step; thus, the half-sample weights are permanently incorporated into the survey records. This procedure has several advantages over the approach of re-computing the factors for each 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. The application of the estimation and sampling error programs, written as any other SAS procedure, is then straight forward and can be accomplished with a minimum of technical assistance.

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. In the next two sections, the first and second stage factors are discussed with references to the formulation given in Appendix A as required.

Substantial secondary data exist for counties making up all of the first stage sampling units (PSU's) in the country. These data are used in ratio estimation to reduce the component of variance contributed by confining the investigation to a sample, rather than to all PSU's.

The numerator of a first stage ratio factor consists of the known group total for all strata in the group. The denominator consists of an estimate of this total obtained by inflating the sampled PSU information to the stratum level (using the inverse of the PSU selection probability), and then summing across all strata in the groups. For a half sample, the denominator is based on data from the PSU's in the half sample. These totals are precomputed and used in the matrix operations shown in Appendix A. In the appendix notation, the known group totals are denoted as matrix A and the estimates of the totals denoted as matrix B.

The second stage of ratio estimation is introduced to reduce the variance component that arises because the accident investigations are conducted for a sample rather than for all accidents within the PSU. The sampling procedure produces a complete listing of all police reported accidents so that, for the PSU, a complete census of all accidents in each of the separate accident category strata is available for the survey period. The numerator of each within PSU ratio factor consists of the known total number of accidents of a specific type weighted up to the total of the stratum from which the PSU was selected. (These figures are stored in a matrix denoted as C in Appendix A). The denominator consists of the estimate of this total based on the sample from the appropriate accident category stratum. (These figures are stored in a matrix denoted as D in the Appendix). Ratios developed in this manner for each of the accident category strata for each NASS PSU would be very unstable due to small numbers of accidents of particular types expected to occur. The matrix routines allow the combining of accidents across PSU strata, however, to form a more stable set of ratios.

In the previous section it was pointed out that primary strata used in the selection of PSU's are collapsed to form between-PSU ratio factors; a similar procedure has been performed for within-PSU ratio factors. The criteria for combining the PSU strata are different so that the PSU groups for the two stages of ratio estimation may not be the same.

4. 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. Variances can be estimated for a wide variety of statistics of interest, linear or non-linear. This paper does not discuss the BRR method in detail; the subject is treated in a number of articles, for example (4) and (5). The programs supplied for the illustration have been written to incorporate several practical problems faced in analyzing the results of a survey 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.

The first step in the application of the BRR method is to define the half samples. Several papers have been written which provide guidance in this task, the illustration employs the method of Plackett and Burman (6). The result of this step is a design matrix defining the half samples. The columns of the matrix identify each half sample (the first column is labelled replicate zero and is used to produce full-sample design. There are two rows for each stratum, one for each of the PSU's in the pair selected from the stratum. This matrix is denoted as "R" in Appendix A.

To more adequately reflect the impact of the estimation procedures on the variance of the estimates, it is necessary to apply the estimator to each of the estimates prepared from the half samples. This involves the preparation of ratio adjustment factors for each half sample. A major stumbling block to the application of such adjustments is the extensive amount of computation required to repeat the computations for each half sample.

Two programs have been written within the SAS system to complete these pre-processing tasks. The first calls on the matrix operations of the SAS procedure, PROC MATRIX. This program performs the matrix operations described in Appendix A using matrices supplied by the user. The second program, a SAS Macro, merges the results of the first program with the survey records.

The use of matrix operations to perform the computations greatly simplifies the preparation of half-sample weights. The steps delineated in Appendix A allow fairly complex rules to be used in the construction of ratio adjustments. Separate variables may be used to adjust different sampling units (post-stratification). In the preparation of ratio factors, strata are pooled; numerators and denominators are prepared within each of these "collapsing groups" of strata. The matrix procedure allows for easy definition of strata to be collapsed; these definitions are realized through the "F" and "G" matrices discussed in the Appendix. If the number of cases in any of the factors is found to be insufficient or if the resulting ratios are too extreme, changes in the "F" or "G" matrices permit easy redefinition of the pooling groups.

5. 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 NASS illustration, 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 OUTVAR statement discussed below. (User specified constants may be included in expressions to allow a third stage of ratio adjustments.)

The annotated listing in Figure 5-1 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, ACCS, is set equal to "1" for all accident records (see line 28).

To obtain an estimate of the total number of urban accidents, a dummy variable, URBAN, is defined in lines 30-32. This variable is a recoding 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 53 and 66) as the ratio of URBAN to ACCS.

The output resulting from invoking NASSTIM (lines 48-54) is shown in Figure 5-2. Three lines are displayed; one for each statistic. The sum of the appropriate weights is given for ACCS 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 60-66 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 5-3.

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 62 provides the names of the weight variables required. Since this example applies to a survey using 10 PSU's (see line 39), five paired strata are necessary with eight replicates to be generated. Thus, nine weights are required; the first weight "R_WGTO", is used to estimate replicate zero, the full sample estimate. Line 66 defines a transformation, the ratio of urban accidents to total accidents.

At the user's option (lines 70-71), the output data set can be printed (Figure5-4). Figure 5-4 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.

Figure 5-1. Example of NASSTIM and NASSVAR usage.

```
NOTE: THE JOB LUCKYSOO HAS BEEN RUN UNDER RELEASE 79.5 OF SAS AT INFORMATICS INC.
                                                                          (00275).
NOTE: SAS OPTIONS SPECIFIED ARE:
      SORT=6

1          /*---------------------------------------------------------*/
2          /*                                                         */
3          /*   EXAMPLE OF RUNNING NASSTIM & NASSVAR                  */
4          /*                                                         */
5          /*   ASSUME THE FOLLOWING ESTIMATES ARE TO BE PRODUCED     */
6          /*   FROM THE 79 ANALYSIS FILE..                           */
7          /*                                                         */
8          /*         1.  NUMBER TOTAL ACCIDENTS (NATIONAL LEVEL)     */
9          /*         2.  NUMBER TOTAL URBAN ACCIDENTS.               */
10         /*         3.  RATIO OF URBAN TO TOTAL ACCIDENTS.          */
11         /*                                                         */
12         /*---------------------------------------------------------*/
13
14         DATA ACCIDENT ;
15              SET INDD1.ACCIDENT ;
16              /*                                                    */
17              /*   READ THE 1979 NASS ANALYSIS FILE.                */
18              /*   -WEIGHT- VARIABLES TO IMPLICITLY DEFINE          */
19              /*   REPLICATE AND RATIO FACTOR HAVE BEEN             */
20              /*   ADDED IN A PREVIOUS STEP.                        */
21              /*                                                    */
22              /*   DELETE TRUCK UNDERRIDE ACCIDENTS FROM ANALYSIS   */
23
24              IF H02 > '4990' THEN DELETE ;
25
26              /*   COUNT ACCIDENTS                                  */
27
28              ACCS=1 ;
29
30             /*   COUNT URBAN ACCIDENTS.                            */
31             IF A21=2 THEN URBAN=1 ;
32             ELSE URBAN=0 ;
33
34             /*   ASSIGN REQUIRED VARIABLE 'NUMPSU' THE VALUE 10     */
35             /*   THIS VARIABLE IS REQUIRED BY NASSVAR TO RESIDE     */
36             /*   ON THE INPUT DATA SET & REPRESENTS THE NUMBER      */
37             /*   OF PSUS IN THE SAMPLE DESIGN BEING ANALYZED.       */
38
39             NUMPSU=10 ;
40
41             LABEL
42             ACCS=ACCIDENTS
43             URBAN=URBAN ACCIDENTS (TOTAL)
44             ;
45
46               /*    INVOKE NASSTIM TO PRODUCE 'CHEAP' ESTIMATES.    */
47

NOTE: DATA SET WORK.ACCIDENT HAS 3331 ORSERVATIONS AND 61 VARIABLES. 55 OBS/TRK.
NOTE: THE DATA STATEMENT USED 1.31 SECONDS AND 184K.

48               PROC NASSTIM DATA=ACCIDENT BEST ;
49                    VAR ACCS URBAN ;
50                    WEIGHT R_WGT0 ;
51                    COMPVAR ACCS URBAN ;
52                    OUTVAR U_RATIO ;
53                        U_RATIO=URBAN / ACCS ;
54               TITLE SELECTED NATIONAL LEVEL ACCIDENT ESTIMATES ;
55
56
57             /*    INVOKE NASSVAR TO PRODUCE ESTIMATES & SAMPLING    */
58             /*    ERRORS.                                           */
59
60

NOTE: NASSTIM IS AN UNSUPPORTED, EXPERIMENTAL PROCEDURE.

WESTAT INC
1650 RESEARCH BLVD
ROCKVILLE,MD 20850
(301) 251-1500

NOTE: THE PROCEDURE NASSTIM USED 1.48 SECONDS AND 184K AND PRINTED PAGE 1.

60             PROC NASSVAR DATA=ACCIDENT BEST TOTAL OUTPUT OUTDATA=TWO ;
61                  VAR ACCS URBAN ;
62                  WEIGHT R_WGT0-R_WGT8 ;
63                  COMPVAR ACCS URBAN ;
64                  OUTVAR U_RATIO ;
65
66                      U_RATIO=URBAN / ACCS ;
67
68             TITLE SELECTED NATIONAL LEVEL ACCIDENT ESTIMATES. ;
69
70

NOTE: NASSVAR IS AN UNSUPPORTED, EXPERIMENTAL PROCEDURE.
NOTE: DATA SET WORK.TWO HAS 9 ORSERVATIONS AND 4 VARIABLES. 361 UBS/TRK.

WESTAT INC
1650 RESEARCH BLVD
ROCKVILLE,MD 20852
(301) 251-1500

NOTE: THE PROCEDURE NASSVAR USED 4.58 SECONDS AND 188K AND PRINTED PAGE 2.

70             PROC PRINT DATA=TWO ;
71             TITLE1 REPLICATE LEVEL ESTIMATES (OUTPUT DS FROM NASSVAR) ;

NOTE: THE PROCEDURE PRINT USED 0.16 SECONDS AND 176K AND PRINTED PAGE 3.

NOTE: SAS USED 188K MEMORY.

NOTE: SAS INSTITUTE INC.
      SAS CIRCLE
      BOX 8000
      CARY, N.C. 27511
```

Figure 5-2. NASSTIM output.

SELECTED NATIONAL LEVEL ACCIDENT ESTIMATES
NASS ESTIMATION PROCEDURE

OPTIONS USED:
BEST

3331 OBSERVATIONS PROCESSED
6704645 WEIGHTED OBSERVATIONS PROCESSED

NAME	LABEL	ESTIMATE	MISSING	WEIGHTED MISSING
ACCS	ACCIDENTS	6704645	0	0
URBAN	URBAN ACCIDENTS (TOTAL)	4674048	0	0
U_RATIO	COMPUTED ESTIMATE	0.697136	N/A	N/A

Figure 5-3. NASSVAR output.

```
OPTIONS USED:
            BEST
            TOTAL
            OUTPUT
       3331 OBSERVATIONS PROCESSED
    6704645 WEIGHTED OBSERVATIONS PROCESSED
          8 REPLICATES IN  10 PSU DESIGN
```

SELECTED NATIONAL LEVEL ACCIDENT ESTIMATES.
NASS SAMPLING ERRORS PROCEDURE

ACCS — ACCIDENTS

ESTIMATE	MISSING	WEIGHTED MISSING	VARIANCE	REL-VARIANCE	CV(%)	STANDARD ERROR	LOWER 95% CONF INTVL	HIGHER 95% CONF INTVL
6704645	0	0	1.742E+11	0.00387527	6.22517	417376	5886589	7522701

URBAN — URBAN ACCIDENTS (TOTAL)

ESTIMATE	MISSING	WEIGHTED MISSING	VARIANCE	REL-VARIANCE	CV(%)	STANDARD ERROR	LOWER 95% CONF INTVL	HIGHER 95% CONF INTVL
4674048	0	0	8.093E+11	0.0370456	19.2472	899625	2910782	6437314

U_RATIO — COMPUTED ESTIMATE

ESTIMATE	MISSING	WEIGHTED MISSING	VARIANCE	REL-VARIANCE	CV(%)	STANDARD ERROR	LOWER 95% CONF INTVL	HIGHER 95% CONF INTVL
0.697136	.	.	0.00972521	0.0200108	14.1459	0.0986165	0.503848	0.890424

Figure 5-4. Listing of NASSVAR output data set

REPLICATE LEVEL ESTIMATES (OUTPUT DS FROM NASSVAR)

OBS	REPL_ID	ACCS	URBAN	U_RATIO
1	0	6704645	4674048	0.697136
2	1	6021238	3154835	0.523951
3	2	6605781	4514185	0.683369
4	3	7269834	5317400	0.731434
5	4	7387759	6134438	0.830352
6	5	6572466	4754515	0.723399
7	6	6478077	3668768	0.566336
8	7	6650030	5263451	0.791493
9	8	6947396	5153569	0.741799

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 "k"). 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.

B_{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.

G_{2hxg} 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 (k+1)x [cx(k+1)] Matrix of 1's and 0's used to collapse over c between-PSU ratio groups:

```
1...10...00.....0...0...0
0...01...10.....0...0...0
:                       :
:                       :
0...00...00....00..01...1
```

I_G (k+1)x [gx(k+1)] Matrix of 1's and 0's used to collapse over g between-PSU ratio groups (see I_F).

$I_{2hx(k+1)}$ 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 required below. These operations are defined as follows:

a. Matrix term-by-term addition:

$$M_{axb} + N_{axb} = 0_{axb}$$

b. Matrix term-by-term multiplication:

$$M_{axb} \# N_{axb} = 0_{axb}$$

c. Matrix transpose:

$$M'_{bxa} = M_{axb}$$

d. Matrix multiplication (dot products):

$$M_{axb} * N_{bxc} = 0_{axc}$$

e. Horizontal direct product:

$$M_{axb} @ N_{axc} = 0_{ax(bxc)}$$

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:

$$M_{axb} \# N_{axb} = 0_{axb}$$

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:

$$J_{2hx[cx(k+1)]} = R_{2hx(k+1)} @ F_{2hxc}$$

b. Compute between-PSU numerator terms:

$$NUM_{ux[cx(k+1)]} = A'_{ux2h} * (I_{2hx(k+1)} @ F_{2hxc})$$

c. Compute between-PSU denominator terms:

$$DEN_{ux[cx(k+1)]} = B'_{ux2h} * J_{2hx[cx(k+1)]}$$

d. Compute between-PSU ratio terms:

$$RATIO_{ux[cx(k+1)]} = NUM_{ux[cx(k+1)]} \# DEN_{ux[cx(k+1)]}$$

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

$$PROD_{[cx(k+1)x(ux2h)]} = RATIO'_{[cx(k+1)]xu} @ SQRT(J')_{[cx(k+1)x2h]}$$

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.

f. Record the between-PSU factor applicable to records identified by replication, PSU, and record type. This step removes redundant zeros generated in previous steps:

$$F_{B(k+1)x(ux2h)} = I_{F(k+1)x[cx(k+1)]} * PROD_{[cx(k+1)]x(ux2h)}$$

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

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

C for A,
D for B,
G for F,
I_G for I_F,
F_W for F_B.

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

3. Compute final ratios as:

$$F_{T(k+1)x(ux2h)} = F_W \# F_B$$

BIBLIOGRAPHY

(1) NASS Estimation - Final Technical Report, National Highway Traffic Safety Administration, Washington, D.C. Contract Number DTNH-80R-07561, July 1982.

(2) National Accident Sampling System (NASS) Final Report Volume I - Final Technical Report, National Highway Traffic Safety Administration, Washington, D.C. Contract number DOT-HS-7-01706, November 1979.

(3) National Accident Sampling System-Estimation, DOT DTNH 22-80R-07561.

(4) McCarthy, Philip J. (1966) "Replication, An Approach to the Analysis of Data from Complex Surveys" Public Health Service Publication No. 1000-Series 2-No. 14.

(5) McCarthy, Philip J. (1969) "Pseudoreplication, Further Evaluation and Application of the Balanced Half-Sample Technique" Public Health Service Publication No. 1000-Series 2-No. 31.

(6) Plackett, R. L. and Burman, J. P., "The Design of Optimum Multifactorial Experiments" Biometrika 33:305-325, 1943-1946.