A SAS® Macro to Calculate Standard Errors for Proportions from Simple Random Sampling within Strata

Wayne E. Johnson, CDC, National Center for Health Statistics

ABSTRACT

Standard statistical programs calculate standard errors as if the data were collected in a simple random sample, yielding incorrect results when applied to data collected according to a complex sample design. Programs which flexibly address complex sample designs and produce appropriate standard errors are available: for example, Shah’s SUDAAN™. Because many data analysts have access to SAS but not to more specialized packages, this paper demonstrates the use of the SAS macro facility to produce proper standard errors for proportions from a common survey design: simple random sampling (with replacement) within strata.

The use of the macro is illustrated with data from the 1988 National Maternal and Infant Health Survey. The example was run upon a '386 IBM®-compatible personal computer with SAS version 6.04 for MS-DOS®.

INTRODUCTION

Simple random sampling within strata means that the survey universe is partitioned into distinct strata and a simple random sample is drawn from each stratum. Surveys based on administrative record systems often follow such a design: using information found in each record, the entire file is sorted into strata; from each stratum a simple random sample is drawn.

In the NMIHS, the system of records consisted of all birth certificates for 1988 in the United States. Strata were formed using the recorded race (Black / Non-Black) and birthweight (L: <1500g, M: 1500-2499g, H: >2500g) to form six strata: BL, BM, BH, NL, NM and NH. From these six strata, simple random samples were drawn with sampling fractions 1/44, 1/55, 1/113, 1/29, 1/160 and 1/720, respectively. Black births and low-weight births were sampled at higher rates to assure adequate sample size. From the form of these sampling fractions, it is easy to see that each birth sampled from stratum BL represents 44 births in the sampling universe; likewise, each birth sampled from stratum BM corresponds to 55 births in the universe, etc. The reciprocals of the sampling fractions are 44, 55, 113, 29, 160 and 720, respectively; these are the initial sampling weights, \( W_{ch} \). Mothers so selected were sent a questionnaire to fill out and mail back; those returned were considered responses. To adjust for non-response, other demographic characteristics such as the mother’s age and marital status, also recorded on the selected birth-certificates, were used to form non-response adjustment cells within strata. In such a cell,
suppose that the number of responses was $R$ and the number of non-responses was $N$. Then the $R$ responses must represent the cell total, $R+N$, so the adjustment factor will be $(R+N)/R$. For each response in the cell, the adjusted weight will be the stratum weight multiplied by $(R+N)/R$: $W_{ih} = W_{0h}*(R+N)/R$. The non-responses were assigned $W_{ih} = 0$ and were henceforth ignored. Thus, observations drawn from the same stratum may have different weights for estimation.

The NMIHS data record for each response contains the variables RACEBWT for the Race/Birthweight class and BASICWGT for the non-response adjusted sampling weight; they convey all the sample design information needed by the macro.

Sometimes we may be interested only in a proper subset of the sample: such a subset is called a domain. This macro requires the user to create a 0/1 variable to indicate which observations belong to the domain (1) and which do not (0).

To illustrate the use of the macro, we shall make three estimates for those mothers who live in metropolitan areas; the list below shows the percentage and the corresponding 0/1 variable:

- age under 19
  _AGELT19 (1=<19, 0=not)

- % married
  _MARRIED (1=married, 0=not)

- % less than 12 yrs education
  _LT12YRS (1=<12 yrs, 0=not)

The variable METRO will set to 1 for those mothers residing in metropolitan areas, 0 otherwise.

For simplicity, the macro expects that the entire input data set will be sorted by the stratum variable; the attributes, for which proportions and standard errors will be estimated, and the domain indicator should be 0/1 variables in which 1 signifies that the case possesses the attribute or belongs to the domain, respectively, and 0 otherwise.

**EXAMPLE**

Suppose that you have already sorted the input data set by stratum and created the necessary indicator variables. On the COMMAND line of the PROGRAM window, enter

```
INCLUDE "C:\SASMACRO\TESTMAC"
```

to bring in the TESTMAC macro:

```
OPTIONS SOURCE2;
FILENAME ASCIIWK1 'C:\ASCIIWK1';
FILENAME ASCIIWK2 'C:\ASCIIWK2';
LIBNAME C 'C:\SASDATA';
%INCLUDE 'C:\SASMACRO\SE1NEW.SAS';
%INCLUDE 'C:\SASMACRO\SE2NEW.SAS';
```

The order of items read within CARDS;...<< is:

1. The data set name, relative to the LIBNAME.
2. The domain indicator.
3. The stratum variable.
4. The weight variable.
5. The attribute variables.

928
To execute the macro, enter SUBMIT upon the COMMAND line of the PROGRAM window. Another macro, SE1NEW.SAS, is invoked automatically; it reads the information in CARDS; ...< and creates global macro variables which are resolved later in the following macro, SE2NEW.SAS. SE2NEW.SAS performs the calculations.

**SE1NEW.SAS**

```sas
OPTIONS SOURCE2 MPRINT MTRACE SYMBOLGEN;
DATA NULL; INFILE CARDS;
RETAIN NAME 0 MINUS1 -1;
LENGTH DATA $17;
LENGTH NAME $17;
RETAIN 'NAME' 'XH' 'YH' 'YH1-YH' 'X1-X' 'Y1-Y' 'YH1-YH' 'WXH' 'WXH1-WXH' 'VH1-VH' 'RH1-RH' 'SUM1-SUM1' 'SUM1-SUM1' 'STDER' 'STDER1-STDER' NAMES 'NAME1-NAMES';
INPUT DATA: SCHAR17.;
CALL SYMPUT('DATA', DATA);
INPUT DOMAIN: SCHAR8.;
CALL SYMPUT('DOMAIN', DOMAIN);
INPUT STRATUM:SCHAR8.;
CALL SYMPUT('STRATUM', STRATUM);
INPUT WEIGHT: SCHAR8.;
CALL SYMPUT('WEIGHT', WEIGHT);
FILE PRINT;
PUT i1 'INPUT SAS DSN:' i35 DATA:$CHAR17.1
i1 'VARIABLE FOR DOMAIN:' i35 DOMAIN:SCHAR8.1
i1 'STRATUM VARIABLE:' i35 STRATUM:SCHAR8.1
i1 'WEIGHTING VARIABLE,' i35 WEIGHT:SCHAR8.1;
LOOP: INPUT i1 VARIABLE:$CHAR8.;
IF (VARIABLE EQ '«') THEN GO TO WRAP_UP;
COUNTER+1:
NC=REVERSE(TRIM(REVERSE(PUT(COUNTER,3.»»;
SUBSTR(NAME,5) = NC;
SUBSTR(VNAME,2) = NC;
FILE ASCIIWK1;
PUT VARIABLE:SB. i;
FILE ASCIIWK2;
PUT +1 NAME:$8.oine VARIABLE:$8. +MINUS1 
+1 $8; 
GO TO LOOP;
WRAP_UP: CALL SYMPUT (‘WV’,COUNTER);
SUBSTR(XH,7) = NC;
CALL SYMPUT (‘XH’,XH);
SUBSTR(YH,7) = NC;
CALL SYMPUT (‘YH’,YH);
SUBSTR(X,5) = NC;
CALL SYMPUT (‘X’,X);
SUBSTR(Y,5) = NC;
```

**SE2NEW.SAS**

```sas
DATA VARIANCE(KEEP=&STRATUM &XH &YH &WXH &WYH &YYH SMALLNH BIGNH)
RATIOS(KEEP=VAR BIGN EX);
RETAIN COUNT1 COUNT2 &X &XH &YH SMALLNH BIGNH EYYNEX 0;
ARRAY X (_VAR_) &X;
ARRAY Y (_VAR_) &Y;
ARRAY R (_VAR_) &R;
ARRAY XH (_VAR_) &XH;
ARRAY YH (_VAR_) &YH;
ARRAY WXH (_VAR_) &WXH;
ARRAY WYH (_VAR_) &WYH;
ARRAY YYH (_VAR_) &YYH;
SET &DATA END;EOF;
BY &STRATUM;
IF (FIRST.&STRATUM)
THEN DO;
SMALLNH = 0;
BIGNH = 0;
DO OVER XH;
XH = 0;
YH = 0;
WXH = 0;
WYH = 0;
YYH = 0;
END;
END;
END;
SMALLNH + 1;
BIGNH + &WEIGHT;
BIGN + &WEIGHT;
COUNT1 + 1;
ARRAY V(_VAR_);
%INCLUDE ASCIIWK1;
SELECT (&DOMAIN);
WHEN (1) DO;
COUNT2 + 1;
DO OVER V;
0 = (V GT .);
IF (0 EQ 1)
THEN DO;
W = &WEIGHT;
WX = W * V;
Y = V * W;
WXH = WX;
WYH = WY;
END;
END;
```
Output from the Macro

The macro echoes back the first part of the user's directions:

```
INPUT SAS DSN: C.NMHS
D/1 VARIABLE FOR DOMAIN: METRO
STRATUM VARIABLE: RACEBWT
WEIGHTING VARIABLE: BASICAGT
```

then it displays

```
TOTAL NUMBER OF OBSERVATIONS
IN INPUT DATASET: 9635
IN DOMAIN: 7835
```

and finally it shows

```
VARIABLE    MEAN       S.E.
  _MARRIED  0.74003   0.00651
  _AGELT19  0.07131   0.00392
  _LT12    0.19817   0.00628
```

The means and standard errors above can be rescaled to percentages by multiplying each figure by 100.

**ALGORITHM**

The means are calculated as the ratio of two weighted sums:

\[
\bar{y} = \frac{\sum (W \cdot y)}{\sum (W \cdot x)}
\]

the numerator is the sum of weights of cases belonging to the domain which possess the attribute and the denominator is the sum of the weights of cases belonging to the domain.

I shall summarize the steps taken to estimate the
variances of these ratios.

(1) For cases which do not belong to the domain of interest, attribute variables and their observed/missing indicators are set to 0. For a justification, see Cochran, pp 35-37.

(2) Cochran, pp 31-32, also shows how the variance of a ratio estimator can be approximated for a simple random sample:

\[ V(\hat{R}) = \frac{1}{(\sum X)^2} V(\sum (y - \hat{R}x)) \]

I have stretched the result to

\[ V(\hat{R}) = \frac{1}{(\sum h=1 \sum i=1 W_{h,i}X_{h,i})^2} \]

\[ V(\sum h=1 \sum i=1 W_{h,i}(Y_{h,i} - \hat{R}X_{h,i})) \]

which simplifies to

\[ V(\hat{R}) = \frac{1}{(\sum h=1 \sum i=1 W_{h,i}X_{h,i})^2} \]

\[ \sum h=1 \sum i=1 V(\sum i=1 Z_{h,i}) \]

where \( Z_{h,i} = W_{h,i}(Y_{h,i} - \hat{R}X_{h,i}) \).

\[ V(\sum i=1 \sum h=1 Z_{h,i}) = \frac{n_h}{n_h-1} \cdot (1 - f_h) \cdot \frac{\sum i=1 Z_{h,i}^2 - n_h \overline{Z}_h^2}{\sum i=1 \sum h=1 W_{h,i}^2 (Y_{h,i} - 2\hat{R}Y_{h,i} + \hat{R}^2)} \]

\[ \overline{Z}_h = \frac{1}{n_h} \sum i=1 \sum h=1 W_{h,i}(Y_{h,i} - \hat{R}X_{h,i}) \]

REFERENCES


ACKNOWLEDGEMENT

This paper grew out of my effort to understand the Taylor linearization performed by SUDAAN; in that sense, my work is completely derivative from SUDAAN. All errors in my work, however, are mine.

I found, for proportions, that my algorithm gave standard errors which agreed closely (to 3 significant digits) with
those produced by SUDAAN. Agreement was poor for means of continuous variables. Thus, I fear that my algorithm does not properly mimic the Taylor linearization performed by SUDAAN, which remains my standard for accuracy.

SAS is a registered trademark of SAS Institute, Inc. in the USA and other countries.

Other brand and product names are registered trademarks or trademarks of their respective companies.