A SAS MACRO FOR ESTIMATING SAMPLING ERRORS OF MEANS AND PROPORTIONS IN THE CONTEXT OF STRATIFIED MULTISTAGE CLUSTER SAMPLING

Mamadou Thiam, Alfredo Aliaga, Macro International, Inc.
Mamadou Thiam, Macro International, Inc. 11785 Beltsville Drive, Suite 300, Calverton, MD 20705

Key Words: Weighting, Stratification, Clustering, Sampling error.

Abstract

In producing estimates from sample surveys, it is essential to quantify the extent of sampling error. Survey data are frequently collected based on complex sample designs featuring unequal weighting, stratification and clustering. Most standard statistical software handle weights correctly in producing point estimates, but few have options to account for the complexity of the sample design in estimating sampling errors. This user-friendly SAS macro allows the computation of sampling error estimates for means and proportions for the total sample as well as specified subclasses in the context of stratified multistage cluster sampling with unequal probabilities of selection. Using the Taylor series approximation to estimate the sampling error from the sample, the macro also produces the unweighted and weighted sample sizes, the sampling error under simple random sampling, the design effect, the relative error and the confidence intervals of the estimate. The coefficient of variation for the size of the primary sampling units is also given to allow a check of the validity of the assumption underlying the Taylor series approximation method.

Introduction

Researchers widely derive estimates of population characteristics using a probability sample drawn from the population under study. Two types of errors affect these estimates: non-sampling errors and sampling errors. Non-sampling errors generally result from mistakes made during the implementation of data collection and data processing. Sampling errors, which we only deal with, arise from limiting the inquiry to a sample of the population. If a probability sample is drawn, the particular sample obtained is only one of the many samples that could have been selected from the same population using the same sample design. Each of these samples would yield estimates that differ somewhat from the estimates derived from the sample that is actually available. The sampling error of an estimate is a measure of the variability between estimates from all possible samples of the same size and design under same essential survey conditions. The degree of this variability can be estimated from the one probability sample that is selected. The computational procedure must take into account the complexity of the sample design, particularly the effects of unequal weighting, stratification, and clustering which influence the extent of the sampling error. Several sampling error estimation programs are currently available, and they all produce similar results when the sample size is large. This paper presents a SAS macro, %STDERROR, that takes into account the complexity of the sample design in estimating the sampling error of means and proportions for the total sample as well as specified subclasses. The macro also produces certain statistics derived from the estimated sampling error.

Computational method and formulae

The %STDERROR macro uses the Taylor linearization method of variance estimation to estimate sampling errors of means and proportions under an ultimate cluster sample selection model. Under the ultimate cluster sampling model, elements within primary sampling units (PSUs) are divided into ultimate clusters and then one ultimate cluster is taken from every PSU. The Taylor linearization method treats any estimator as a ratio estimator, obtains a linear approximation of the estimator, and then uses the variance of the approximation to estimate the variance of the estimator itself. This methods also assumes that two or more primary sampling units (PSUs) are selected independently and with replacement from each stratum. In practice, PSUs are usually selected without replacement and variance estimates obtained using the Taylor linearization method are then overstated, but this should be negligible if the sampling fraction for PSUs is small. With the Taylor linearization method, aggregated quantities are computed only at the PSU level; the design information at other subsequent stages is not used. It is worth noting that the strata used for computing sampling errors are not necessarily the explicit strata used in sample design and selection. In the case they differ, the original explicit strata are of no interest. Strata to be used for the computation of sampling errors may be created by grouping PSUs based on some characteristic of similarity so as to maintain homogeneity within each stratum.

To supply the basic formulae, we consider a ratio estimator r = y/x where y and x are weighted sums over the whole sample or a subclass. Suppose there are H strata in the sample or subclass, and m_h PSUs are selected from stratum h. For any PSU i in stratum h, the notations follow:

y_{hij} = value of variable y for element j in PSU i in stratum h;

y_{hi} = weighted sum of all values y_{hij} for all elements in PSU i, i.e.,

$$y_{hi} = \sum_{j=1}^{m_{hi}} w_{hij}\, y_{hij}$$

where w_{hij} is the sample weight for element j in PSU i in stratum h, and m_{hi} is the number of elements in PSU i.

y_h= sum of values for stratum h, i.e.,

$$y_h = \sum_{i=1}^{m_h} y_{hi}$$

y = sum over the whole sample or subclass, i.e.,

$$y = \sum_{h=1}^{H} y_h$$

Similar notations apply for variable x.

The variance of the ratio estimator r is computed using the formula given below, with the sampling error being the square root of the variance:

$$SE^2(r) = var(r) = \frac{1-f}{x^2} \sum_{h=1}^{H} \left[\frac{m_h}{m_h - 1} \left(\sum_{i=1}^{m_h} z_{hi}^2 - \frac{z_h^2}{m_h} \right) \right]$$

in which $z_{hi} = y_{hi} - r.x_{hi}$, and $z_h = y_h - r.x_h$

where f is the overall sampling fraction, which is so small that it can be ignored.

The estimate of the sampling error given by the Taylor linearization method is not unbiased. The magnitude of this bias depends on the coefficient of variation of primary sampling units (CV) and may be ignored for variation less than 0.2. %STDERROR prints the coefficient of variation of PSU sizes for the entire sample as well as for every subclass. This CV is computed using the following formula:

$$CV = \frac{1}{x} \left[\sum_{h=1}^{H} \frac{m_h}{m_h - 1} \left(\sum_{i=1}^{m_{hi}} x_{hi}^2 - \frac{\left[\sum_{i=1}^{m_{hi}} x_{hi} \right]^2}{m_h} \right) \right]^{\frac{1}{2}}$$

where x_{hi} is the weighted size (weighted number of elements) for PSU i.

The design effect, DEFT, printed by %STDERROR is the ratio of the estimated sampling error under the actual sample design to the sampling error that would be obtained under simple random sampling. Assuming simple random sampling, the sampling error for the ratio estimator r = y/x is given by the following formula:

$$SE_{srs}^2\ (r) = \frac{1-f}{n-1} \; \frac{\sum_{h=1}^{H} \sum_{i=1}^{m_h} \sum_{j=1}^{m_{hi}} w_{hij}\, z_{hij}^2}{\sum_{h=1}^{H} \sum_{i=1}^{m_h} \sum_{j=1}^{m_{hi}} w_{hij}}$$

where $z_{hij} = y_{hij} - r.x_{hij}$

Syntax and Input

The following statement, in which key parameters are to be specified, will invoke and execute the %STDERROR macro within a SAS program:

```
%STDERROR (DATA=, WEIGHT=, PSU=,
STRATA=, BYVAR=, ALLVARS=);
```

In the above macro call, the DATA= statement names the input data set which should contain all variables for which the sampling error is requested, and any other variables identified in the WEIGHT, PSU, STRATA, and BYVAR statements. The WEIGHT= statement identifies the weight variable. The PSU= statement specifies the primary sampling unit to which each ultimate cluster in the sample belongs whereas STRATA= names the stratification variable to be used for estimating sampling errors. If sampling errors are to be estimated for subclasses, the BYVAR= statement must name the variable defining the subclasses. All variables for which sampling errors are to be computed, should be listed and separated with a space in the ALLVARS= statement. It should be noted that except the BYVAR= statement which is optional, all other statements are required for the macro to work properly. If subclass estimates are requested, then the macro will produce estimates for all subclasses as well as for the whole sample.

Output

%STDERROR creates and prints a SAS data set containing the following results for every variable for which the sampling error was requested:

CV : Coefficient of variation of PSU sizes

VARIABLE : Variable name

LABEL : Variable label, if any. The maximum length is 40 characters including spaces.

TYPE : Type of statistics (mean or proportion)

MEAN : Weighted sample mean or proportion

STDERROR : Sampling error of the sample mean or proportion

N_UNWGT : Unweighted number of cases used in the calculations

N_WGT : Weighted number of cases used in the calculations

SRS : Sampling error of the mean or proportion under simple random sampling

DEFT : Design effect

RELERROR : Relative error of the sample mean or proportion

LOWER : Lower bound of the 95% confidence intervals

UPPER : Upper bound of the 95% confidence intervals

The resulting estimates were compared with estimates obtained through the Taylor linearization method used in SUDAAN and the sampling error module of the Integrated System for Survey Analysis (ISSA). All three programs produced the same estimates.

Examples

The examples in this paper use the data set for a country in which a Demographic and Health Survey (DHS) was conducted. The survey objectives include the estimation of the prevalence of contraceptive use in women 15-49 years old, and of fertility and childhood mortality rates. For the purpose of the illustration, the following selected variables are used:

V102 : type of residence (urban, rural)
V106 : highest education level attended
V201 : total number of children ever born to the woman
V502 : marital status
V613 : ideal number of children

A sample of 7060 women with completed interviews was obtained using a stratified two-stage design. In the first stage, 270 PSUs were selected. A complete listing of households within selected PSUs was carried out. The lists of households obtained served as sampling frame for the selection of households in the second stage. All eligible women identified in the selected households were included in the sample. The sample take within each PSU corresponds to the ultimate cluster.

Categorical variables need to be recoded in order to estimate the proportions that are of interest to the survey. The following statements create the data set TEST that will contain the new variables to be used in the macro call.

```
    /***** RECODING VARIABLES   ******/

DATA TEST; SET MDIR;
 RWEIGHT=V005/1000000;     /*weight variable */
 EVBORN = V201;
 IF V102=1 THEN URBAN=1; ELSE URBAN=0;
 IF V106 IN (2,3) THEN EDUC=1; ELSE EDUC=0;
 IF 0<=V613<=45 THEN IDEAL=V613; ELSE   IDEAL=.;
 IF V502=1 THEN CURMAR=1; ELSE CURMAR=0;
  LABEL URBAN    = 'Urban residence'
           EDUC      = 'With secondary education or higher'
          CURMAR  = 'Currently married (in union)'
          EVBORN   = 'Children ever born'
           IDEAL      = 'Ideal number of children';
RUN;
```

Example 1: Sampling errors for subclasses are not requested.

```
%INCLUDE 'C:\MEANS\STDERROR.SAS';

%STDERROR (DATA= test, WEIGHT=  rweight,
PSU=v021, STRATA=v022,  ALLVARS= urban
educ curmar evborn ideal);
```

In this example, the %include statement brings into the SAS program the %stderror macro which is stored in the directory 'c:\means'. Sampling errors will be produced when the macro call statement %stderror is executed.

Example 2: Sampling errors for subclasses are requested.

```
%INCLUDE 'C:\MEANS\STDERROR.SAS';

%STDERROR (DATA=test, WEIGHT= rweight,
PSU=v021, STRATA=v022, BYVAR=v102,
ALLVARS=urban educ curmar evborn ideal);
```

The output files for these examples are shown in annex.

Acknowledgments

We are grateful to Pedro Saavedra for his comments and suggestions.

References

Verma, V., Pearce, M. (1986), Clusters, A Package Program for the Computation of Sampling Errors for Clustered Samples, Version 3.0

DHS-III Basic Documentation Number 6 (1996), Sampling Manual.

Authors

Mamadou Thiam, Demographic and Health Surveys program, Macro International, Inc.,
11785 Beltsville Drive, Calverton, MD 20705.
Phone (301) 572 0495, Fax (301) 0572 0999.
E-mail: mthiam@macroint.com

Alfredo Aliaga, Demographic and Health Surveys program, Macro International, Inc.,
11785 Beltsville Drive, Calverton, MD 20705.
Phone (301) 572 0940, Fax (301) 0572 0999.
E-mail: aliaga@macroint.com

Output 1: Sampling errors for subclasses are not requested

```
                                COEFFICIENT OF VARIATION FOR CLUSTER SIZES: ENTIRE SAMPLE

                                                        VARIABLE      CV

                                                         CURMAR       0.029
                                                         EDUC         0.029
                                                         EVBORN       0.029
                                                         IDEAL        0.029
                                                         URBAN        0.029

                                SAMPLING ERRORS: ENTIRE SAMPLE

VARIABLE    LABEL                              TYPE        MEAN   STDERROR   N_UNWGT    N_WGT   SRS    DEFT   RELERROR    LOWER   UPPER

CURMAR      Currently married (in union)       Proportion  0.628  0.008      7060       7060    0.006  1.328  0.012       0.613   0.643
EDUC        With secondary education or higher Proportion  0.269  0.012      7060       7060    0.005  2.310  0.045       0.244   0.293
EVBORN      Children ever born                 Mean        3.215  0.050      7060       7060    0.038  1.308  0.016       3.115   3.315
IDEAL       Ideal number of children           Mean        5.306  0.082      6447       6358    0.035  2.303  0.015       5.142   5.469
URBAN       Urban residence                    Proportion  0.281  0.011      7060       7060    0.005  2.007  0.038       0.260   0.303
```

Output 2: Sampling errors for subclasses are requested

```
                                COEFFICIENT OF VARIATION FOR CLUSTER SIZES: ENTIRE SAMPLE

                                                        VARIABLE      CV

                                                        CURMAR        0.029
                                                        EDUC          0.029
                                                        EVBORN        0.029
                                                        IDEAL         0.029
                                                        URBAN         0.029

                                SAMPLING ERRORS: ENTIRE SAMPLE

VARIABLE    LABEL                              TYPE        MEAN   STDERROR   N_UNWGT    N_WGT   SRS    DEFT   RELERROR    LOWER   UPPER

CURMAR      Currently married (in union)       Proportion  0.628  0.008      7060       7060    0.006  1.328  0.012       0.613   0.643
EDUC        With secondary education or higher Proportion  0.269  0.012      7060       7060    0.005  2.310  0.045       0.244   0.293
EVBORN      Children ever born                 Mean        3.215  0.050      7060       7060    0.038  1.308  0.016       3.115   3.315
IDEAL       Ideal number of children           Mean        5.306  0.082      6447       6358    0.035  2.303  0.015       5.142   5.469
URBAN       Urban residence                    Proportion  0.281  0.011      7060       7060    0.005  2.007  0.038       0.260   0.303
```

```
COEFFICIENT OF VARIATION FOR SAMPLING UNIT SIZES: SUBCLASSES

                    ------------------------------------------------------ V102=1 ------------------------------------------------------

                                       VARIABLE        CV

                                       CURMAR         0.038
                                       EDUC           0.038
                                       EVBORN         0.038
                                       IDEAL          0.039
                                       URBAN          0.038

                    ------------------------------------------------------ V102=2 ------------------------------------------------------

                                       VARIABLE        CV

                                       CURMAR         0.037
                                       EDUC           0.037
                                       EVBORN         0.037
                                       IDEAL          0.037
                                       URBAN          0.037

SAMPLING ERRORS:  SUBCLASSES

                    ------------------------------------------------------ V102=1 ------------------------------------------------------

VARIABLE   LABEL                               TYPE         MEAN   STDERROR  N_UNWGT  N_WGT  SRS    DEFT   RELERROR  LOWER  UPPER

 CURMAR    Currently married (in union)        Proportion   0.570  0.015     2376     1987   0.010  1.450  0.026     0.540  0.599
 EDUC      With secondary education or higher  Proportion   0.526  0.031     2376     1987   0.010  3.043  0.059     0.463  0.588
 EVBORN    Children ever born                  Mean         2.496  0.070     2376     1987   0.057  1.225  0.028     2.356  2.636
 IDEAL     Ideal number of children            Mean         4.181  0.115     2255     1857   0.046  2.469  0.027     3.951  4.410
 URBAN     Urban residence                     Proportion   1.000  0.000     2376     1987   0.000  -------  0.000   1.000  1.000

                    ------------------------------------------------------ V102=2 ------------------------------------------------------

VARIABLE   LABEL                               TYPE         MEAN   STDERROR  N_UNWGT  N_WGT  SRS    DEFT   RELERROR  LOWER  UPPER

CURMAR     Currently married (in union)        Proportion   0.651  0.009     4684     5073   0.007  1.314  0.014     0.633  0.669
EDUC       With secondary education or higher  Proportion   0.168  0.010     4684     5073   0.005  1.884  0.061     0.147  0.189
EVBORN     Children ever born                  Mean         3.496  0.061     4684     5073   0.048  1.260  0.017     3.374  3.619
IDEAL      Ideal number of children            Mean         5.770  0.105     4192     4501   0.046  2.306  0.018     5.560  5.980
URBAN      Urban residence                     Proportion   0.000  0.000     4684     5073   0.000  -------  -------  0.000  0.000
```