PROC SURVEYMEANS

Roger Goodwin

Abstract

This presentation shows how to summarize "closed and weighted" estimates of the area frame data using PROC SURVEYMEANS in SAS®. We compared results from PROC SURVEYMEANS with results from the survey summary package used by the agency. Results were surprisingly close. Since the list frame uses similar estimators, this presentation can be extended to estimate those items also. However, the expansion factors will be inherently different. In the area frame, expansion factors are based on segments, whereas, in the list frame, expansion factors are based on farm operators. For documentation on extremely complex SAS software for calculating weights, see Goodwin (2001, 2001, 2002).

The stratified estimators and the rationale behind using them have been documented for use in the June Surveys (Houseman, 1975; Kott, 1988; Kott, 1990). The SAS macro %QAS is used to calculate totals and standard errors at NASS. It has been modified several times. In that regard, several versions exist. Some versions calculate the standard errors correctly, while other versions do not. Such discrepancies have a significant impact on the national allocation program. For a detailed description of the national allocation program, see Bethel (1986) and Mergerson (1989). This presentation shows a simple way to estimate population sums, variances, and CVs as opposed to using complex data step programming.

1.0 Introduction

The Area Frame Section at USDA/NASS uses a stratified survey design to obtain population totals, standard errors, and CVs. The June Area Survey is conduced every year to measure, among other things, corn acreage, cotton acreage, soybean acreage, three seasons of wheat acreage, NOL cattle, and number of farms in the United States. Each state has a set of land use strata specific to it. Within each strata, replicates are assigned to sub-strata. A sample of land called a *segment* is finally sampled. Since a minimum of two replicates is assigned to the sub-strata, the minimum number of segments per strata is two (assuming one sub-strata). Segment sizes range from $1/10^{th}$ of a square mile to 1 square mile (640 acres per mile). Often, multiple sub-strata are assigned to each stratum.

A complete, comprehensive set of designs for the United States can be found in the *2005 Area Frame Design* book for the 2005 June Area Survey.

2.0 Estimators

One of the output files from the June Area Survey is the segment file. This file is structured such that there exists one observation for each segment. Multiple variables exist for each segment. The closed and weighted estimators of the area frame are listed in Table 1.

Table 1: Estimators

Total $\hat{T}$	**Variance of** $\hat{T}$	**CV of** $\hat{T}$
$\sum_{h=1}^{H}\sum_{i=1}^{n_h}\frac{N_h}{n_h}Y_{hi}$	$\sum_{h=1}^{H} Var(N_h\overline{Y}_h) = \sum_{h=1}^{H}\frac{N_h^2 s_h^2}{n_h}$	$\frac{\sqrt{Var(\hat{T})}}{\hat{T}}$

N_h is the total number of segments in stratum h, and n_h is the number of segments sampled. Y_{hi} is the i-th observation in stratum h for the variable Y. s_h^2 is the usual random part to the variance. As noted in the previous section, there are eight variables of interest. Note that the variance of $\hat{T}$ is based on the average $\overline{Y}_h$, not the total Y_h.

3.0 Estimation

To estimate the crop acreages, NOL cattle, and number of farms, the survey design book (a SAS dataset) must be merged onto the segment data file (another SAS dataset). It is a simple matched merge by state, strata, and sub-strata. Calculate the strata expansion factors, N_h/n_h, in the same data step. The population estimates, standard errors, and CVs are calculated for the United States (National), for each state, and for each stratum in the states. This requires three calls to PROC SURVEYMEANS and changing the BY statement. See Table 2 for the syntax to PROC SURVEYMEANS and some examples.

Table 2: Syntax and Examples

Syntax	**PROC SURVEYMEANS** < options > < statistic-keywords > **;** **BY** variables **;** **CLASS** variables **;** **CLUSTER** variables **;** **DOMAIN** variables < variable*variable*variable ... > **;** **RATIO** < 'label' > variables / variables **;** **STRATA** variables < / option > **;** **VAR** variables **;** **WEIGHT** variable **;**
National Estimates	**PROC SURVEYMEANS** data = segmball04 sum cvsum **;** **VAR** &crops**;** **STRATA** state stratax substrat**;** **WEIGHT** mexpfctr **;** **TITLE** "National Totals, Standard Errors, and CVs"; **Run;**
State Estimates	**PROC SURVEYMEANS** data = segmball04 sum cvsum **;** **BY** state **;** **VAR** &crops**;** **STRATA** state stratax substrat**;** **WEIGHT** mexpfctr **;** **TITLE** "State Totals, Standard Errors, and CVs"; **Run;**

Strata Estimates within State	**PROC SURVEYMEANS** data = segmball04 sum cvsum **;** **BY** state stratax**;** **VAR** &crops**;** **STRATA** state stratax substrat**;** **WEIGHT** mexpfctr **;** **TITLE1** "State Totals, Standard Errors, and CVs"; **TITLE2** "By Strata"; **Run;**

Some notes:

- The options SUM and CVSUM limit the output statistics
- The BY statement becomes more detailed for more finer geography.
- The macro variable &CROPS contains the list of eight variables listed earlier.
- The variable MEXPFCTR contains the weights N_h/n_h.
- TITLE and LABEL statements can be used for more readable output.
- By default, the output statistics go to the SAS Output Window.

Although NASS only publishes National and State totals, the standard errors and CVs are used for planning purposes for the next June Area Survey.

Saving the output statistics for analytic purposes requires three lines of SAS ODS code.

3.1 An Example

Consider the following hypothetical data in Table 3.

Table 3: A Numerical Example

Stratum	Y	MEXPFCTR	Statistics
A	1.2	10	$\hat{T}_A = 169$,
A	7.5	10	$s_A^2 = 14.863333\ldots$
A	8.2	10	$Var(\hat{T}_A) = 4458.9999\ldots$
B	5.4	14	$\hat{T}_B = 90.3$,
B	0.05	14	$s_B^2 = 8.1475$,
B	1.0	14	$Var(\hat{T}_B) = 4790.73$.
C	9.7	9	$\hat{T}_C = 161.1$,
C	8.2	9	$s_C^2 = 1.125$,
			$Var(\hat{T}_C) = 182.25$.
Population Estimates			$\hat{T} = 420.4$,
			$Var(\hat{T}) = 9431.97999$,
			$SE(\hat{T}) = 97.118381$,
			$CV(\hat{T}) = 0.231014$

The following code calculates the *population* estimates, standard errors, and CVs.

```
/* read the hypothetical data into a SAS data set */

data example;
input strata $ ctn mexpfctr;
cards;
a   1.2   10
a   7.5   10
a   8.2   10
b   5.4   14
b   0.05  14
b   1.0   14
c   9.7    9
c   8.2    9;
run;

/* ods code to save the output statistics */
```

```
ods trace on;
ods output statistics=state_stats;

/* calculate the population estimates, standard errors and CVs */

proc surveymeans data = example sum cvsum;
class strata;
strata  strata;
var ctn;
weight mexpfctr;
label strata = "Strata";
label ctn = "Cotton";
label mexpfctr = "Expansion Factor";
title "An Example of PROC SURVEYMEANS";
run;


ods trace off; /* turn ods off */
```

An Example of PROC SURVEYMEANS 1

The SURVEYMEANS Procedure Data Summary

Number of Strata	3
Number of Observations	8
Sum of Weights	90

Statistics

Variable	Label	Sum	Std Dev	Coeff of Variation for Sum
ctn	Cotton	420.400000	97.118381	0.231014

Using a calculator, it is easy to verify that PROC SURVEYMEANS calculates the estimates, standard errors, and CVs correctly. *How did SAS determine* N_h? It is needed to calculate the standard errors. One way to do it is to count the number of observations in each stratum to give n_h. Then, deduce that N_h = MEXPFCTR*n_h.

The three lines of ODS code will produce a SAS data set with three observations (add the BY STRATA statement also). See Table 4. Note the naming convention of the output statistics. It is always the variable name followed by _SUM for totals, _CVSUM for CVs and _StdDev for standard errors.

Table 4: ODS Output File

STRATA	CTN	LABEL_CTN	CTN_SUM	CTN_CVSUM	CTN_StdDev
A	CTN	Cotton	169.000000	0.395123	66.775744
B	CTN	Cotton	90.300000	0.766502	69.215100
C	CTN	Cotton	161.100000	0.083799	13.500000

For multiple variables, the output statistics will appear as new columns next to the existing ones.

4.0 Some PROC SURVEYSELECT Notes

1. User's who use PROC SURVEYSELECT will automatically have the expansion factor variable named "SAMPLINGWEIGHT" on their output file. Use PROC SURVEYMEANS with that variable on the WEIGHT statement.

2. For jackknife estimation, user's who use PROC SURVEYSELECT will automatically have the replicate variable named "REPLICATE" on their output file. Use PROC SURVEYMEANS with that variable on the BY statement.

5.0 References

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

[1] Bethel, James, "An Optimum Allocation Algorithm for Multivariate Surveys," SRS Staff Report, USDA, SF&SRB-98, January 1986.

[2] Goodwin, Roger L., "%ADJWGT_S User's Guide," US Census Bureau, Economic Directorate, Standard Economic Processing System, April 2001.

[3] Goodwin, Roger L., "%WGTSET User's Guide," US Census Bureau, Economic Directorate, Standard Economic Processing System, April 2001.

[4] Goodwin, Roger L., "%ADJWGT_SR User's Guide," US Census Bureau, Economic Directorate, Standard Economic Processing System, January 2002.

[5] Houseman, Earl E., "Area Frame Sampling in Agriculture," Statistical Reporting Service, USDA, SRS No. 20, November, 1975.

[6] Kott, Phillip, "Estimating Variances for the June Enumerative Survey," SRB Staff Report, USDA, SRB-88-06, May, 1988.

[7] Kott, Phillip, “Mathematical Formulae for the 1989 Survey Processing System (SPS) Summary,” NASS Staff Report, USDA, SRB-90-08, May 1990.

[8] Mergerson, James W., “National Area Sample Allocation Analysis,” NASS Staff Report, USDA, SSB-98-01, March, 1989.

[9] SAS Institute Inc., *SAS/STAT User’s Guide, Version 8*, Cary, NC: SAS Institute Inc, 1999. 3884 pp.

6.0 Contact Information

Roger L. Goodwin
U.S. Department of Agriculture
National Agricultural Statistics Service
Area Frame Section
3251 Old Lee Highway
Suite 301
Fairfax, VA 22030
Roger_Goodwin@nass.usda.gov
703-877-8000, ext 120