Journal of Modern Applied Statistical Methods

Volume 15 | Issue 1 Article 48

5-1-2016

Analyzing Different Sampling Designs (SAS)

Ying Lu
Educational Testing Service, ylu@ets.org

Analyzing Different Sampling Designs (SAS)

Cover Page Footnote
The author would like to thank Professor John P. Buonaccorsi for his valuable input, advice and suggestions.

Journal of Modern Applied Statistical Methods
May 2016, Vol. 15, No. 1, 914-956.

ISSN 1538 − 9472

Statistical Software Applications & Review
Analyzing Different Sampling Designs (SAS)

Ying Lu
Educational Testing Service
Rosedale Road, NJ

Various sampling designs are reviewed within the framework of probability sampling. SAS® code to estimate means and proportions, and their standard errors, using different sampling designs are illustrated using example data sets.

Keywords: Sampling, SAS

Introduction

Researchers and statisticians often find it necessary to apply survey-sampling methodologies to acquire information about a large population. Sampling can take different forms (e.g., simple random sampling, stratified sampling, or clustering sampling) and different levels. In order to make appropriate and statistically valid inferences about the population based on the selected sample, the sampling design needs to be taken into consideration in the data analysis.

The purpose of this article is to provide step-by-step guidance on how data obtained from various sampling designs could be using SAS (ver. 9.2) PROC SURVEYMEANS, and to illustrate how to estimate means and proportions, and their standard errors, in various finite sampling designs within the framework of probability sampling. In situations when it is less straightforward to use PROC SURVEYMEANS to obtain the estimates, the use of PROC IML as an alternative tool is illustrated. SAS/IML is an interface that provides interactive matrix programming. It is a separate component from SAS that may require additional installation.

For each sampling design listed below, a brief summary of the model and procedure including formulas for the estimated statistics, their variances, and approximate confidence intervals will be presented. The sampling designs

Ying Lu is a Senior Psychometrician. Email her at: ylu@ets.org.

discussed in this paper all belong to the category of sampling without replacement. Sampling-with-replacement designs are comparatively easy to analyze and therefore are not discussed here. More details and references for the sampling models, designs, and proofs for the formulas used in this paper, as well as definitions and terms adopted in this paper may be found in Lohr (1999). SAS features are demonstrated using example data sets, SAS programs, and output. Data sets used in the examples are selected from the CD accompanying the Lohr book so that interested readers can have access to them. This paper assumes that nonsampling errors such as selection bias and inaccuracy of responses can be ignored in the sampling designs. The sampling designs discussed in this paper are as follows:

- simple random sampling,
- stratified sampling with a Simple Random Sample (SRS) selected from each stratum,
- one-stage cluster sampling with an SRS of clusters,
- two-stage cluster sampling with an SRS at each stage,
- stratified sampling with one-stage cluster sampling (using SRS) within each stratum,
- one-stage cluster sampling with unequal probabilities,
- general complex surveys.

The notations used in this paper differ by section and sampling design. In the more general setting, consider U to be a finite population, and S to be a selected sample. Within the complex sampling framework, subscripts are added to S to denote the sample within a specific cluster or stratum.

Methodology

Simple Random Sampling

Estimating population mean and total Let y_i be the value of interest associated with the i^{th} unit in the population, and let $\bar{y}$ and s^2 denote the sample mean and sample variance, respectively. The population mean $\bar{y}_U$ in an SRS is estimated by the sample mean

$$\hat{\bar{y}}_U = \bar{y} = \frac{1}{n}\sum_{i \in S} y_i \qquad (1)$$

with variance reported as

$$\hat{\mathrm{V}}\left(\hat{\bar{y}}\right)=\left(1-\frac{n}{N}\right)\frac{s^2}{n} \tag{2}$$

and the population total is estimated by

$$\hat{t}=N\bar{y} \tag{3}$$

with variance estimated by

$$\hat{\mathrm{V}}\left(\hat{t}\right)=N^2\left(1-\frac{n}{N}\right)\frac{s^2}{n}\ . \tag{4}$$

For estimating the mean, PROC SURVEYMEANS can be run using only the option of 'total = *N*', which specifies the population size to enable a finite population correction. There are two ways of estimating the total. A weight variable of *N*/*n* can be created, and the sum option can be used. Notice that a common weight on all variables will not affect the estimation of the mean. Alternatively, a new variable can be created that equals *N***y*, where *y* is the original variable and the mean can be computed on this variable.

As an example, 'counties.dat' (Lohr, 1999, p. 440) is used to illustrate the estimation of the population mean and total. The data set contains information on land area, population, numbers of physicians, unemployment, and a number of other quantities for an SRS of 100 counties from the 3,141 counties in the United States. The mean and number of physicians are estimated, along with their associated standard errors and 95% confidence intervals. The SAS code used to obtain the estimates is provided below together with the output.

SAS code:

```
data counties;
infile 'C:\Sampling\counties.dat' dlm=',' firstobs=2;
input RN STATE $ COUNTY $ LANDAREA TOTPOP PHYSICIA ENROLL PERCPUB
CIVLABOR UNEMP FARMPOP NUMFARM FARMACRE FEDGRANT FEDCIV MILIT VETERANS
PERCVIET;
```

```
 /* convert missing values */
  array numval[*] _numeric_;
  do i = 1 to dim(numval);
  if numval[i]=-99 then numval[i]=.;
  end;
  drop i;
  run;

data srs;
set counties;
wt=3141/100;
run;

proc surveymeans data=srs mean clm sum clsum total=3141;
var PHYSICIA;
weight wt;
run;
```

The first part of the SAS code reads in the data and converts the missing value coding of '–99' to the SAS default coding of a period. The second part of the code defines the weight variable as *N*/*n*. The SURVEYMEANS procedure is in the third part of the code with the *clm* and *clsum* options added to give confidence intervals as well as estimates for the mean and the total. The output of the program is displayed below:

```
                    The SURVEYMEANS Procedure

                          Data Summary

                 Number of Observations      100
                 Sum of Weights             3141


                          Statistics

                      Std Error    Lower 95%    Upper 95%
Variable        Mean     of Mean  CL for Mean  CL for Mean     Sum  Std Dev
PHYSICIA  297.170000  156.632533   -13.622928   607.962928  933411   491983


                          Statistics

                               Lower 95%    Upper 95%
                    Variable  CL for Sum   CL for Sum
                    PHYSICIA      -42790      1909612
```

The SURVEYMEANS procedure produces estimated mean, the standard error (SE) of mean, and confidence intervals. The estimated population mean is $\hat{\bar{y}}_U = 297.17$ with $\mathrm{SE}\left(\hat{\bar{y}}\right) = 156.6325$. The estimated population total is $\hat{t} = 933,411$ with $\mathrm{SE}\left(\hat{t}\right) = 491,983$.

Estimating the proportion As a special case of mean estimation, the population proportion *p* in an SRS is estimated by the sample proportion

$$\hat{p} = \bar{y} \tag{5}$$

and

$$\mathrm{SE}\left(\hat{p}\right) = \sqrt{\left(1 - \frac{n}{N}\right)\frac{\hat{p}\left(1-\hat{p}\right)}{n-1}} \tag{6}$$

For the estimation of the proportion, a variable needs to be created so that it takes the value of 1 if the unit has the characteristic of interest and 0 otherwise, and then PROC SURVEYMEANS can be run on the new variable. In the example below, we are interested in the percentage of children that are overdue for a vaccination in a school. Suppose the population consists of 120 children, and the selected sample consists of 10 children with 4 of them being overdue.

```
data a;
 input patient status $ @@;
 cards;
 1 ok 2 ok 3 ok 4 overdue 5 overdue
 6 ok 7 overdue 8 ok 9 overdue 10 ok
 ;;
run;

data a;
 set a;
 if status='overdue' then y=1;
 else y=0;
 wt=120/10;
 run;

proc surveymeans data=a mean clm total=120;
var y;
weight wt;
```

```
run;
```

```
                    The SURVEYMEANS Procedure

                          Data Summary

                  Number of Observations      10
                  Sum of Weights             120


                          Statistics

                              Std Error     Lower 95%     Upper 95%
      Variable      Mean     of Mean    CL for Mean   CL for Mean
      y         0.400000    0.156347      0.046318      0.753682
```

Therefore the estimated proportion is 0.4000 with the SE being 0.1563.

Ratio Estimation

The use of ratio estimation requires measures of y_i and x_i on each sampling unit. The ratio of the two quantities is defined as

$$B = \frac{\overline{y}_U}{x_U} \ . \tag{7}$$

It is estimated by

$$\hat{B} = \frac{\overline{y}}{\overline{x}} = \frac{\hat{t}_y}{\hat{t}_x} \tag{8}$$

with estimated variance

$$\hat{\mathrm{V}}\left[\hat{B}\right] = \left(1-\frac{n}{N}\right)\frac{1}{n\bar{x}_U^2}\frac{\sum_{i\in S}\left(yi-\hat{B}x_i\right)^2}{n-1} = \left(1-\frac{n}{N}\right)\frac{s_e^2}{n\bar{x}_U^2} \tag{9}$$

where $e_i = y_i - \hat{B}x_i$. When $\bar{x}_U$ is not known, we use $\bar{x}$, the sample mean, to approximate it.

Ratio estimation may not be run directly using PROC SURVEYMEANS. Two procedures are demonstrated here, using PROC SURVEYMEANS after some preliminary analyses that feed into it, or via the PROC IML procedure.

For PROC SURVEYMEANS to obtain the correct standard error for ratio estimation, we need to first create a new variable $d = \frac{y-\hat{B}x}{\bar{x}_U}$, and run PROC SURVEYMEANS on *d* as in SRS. As an example, we use 'counties.dat' to estimate the average farm population per square mile of land area. The *y* variable is the farm population and the *x* variable is the land area. Here, $\bar{x}_U$ is assumed to be unknown, and therefore is approximated by $\bar{x}$. The SAS program applied is as follows:

```
proc means mean data=counties noway;
var FARMPOP LANDAREA;
run;

data ratio;
set counties;
d=(FARMPOP-LANDAREA*1.2137218)/944.92;
       /* 944.92 is the sample mean of x obtained from proc means.
         1.213718 is the ratio estimate computed from proc means
         output. */
run;

proc surveymeans data=ratio total=3141;
var d;
run;
```

The first part of the program uses the MEANS procedure to get the sample means of *x* and *y*. The ratio estimate would be computed as $\bar{y}/\bar{x}$. Next the program creates the variable *d*, which is computed from the output we obtained from running the first part of the code. And lastly PROC SURVEYMEANS for SRS is run on the

variable *d*. The resulting SE for the mean of *d*, 0.1891, is the desired SE for the ratio estimate.

The MEANS Procedure

Variable	Mean
FARMPOP	1146.87
LANDAREA	944.9200000

The SURVEYMEANS Procedure

Data Summary

Number of Observations 100

Statistics

Variable	N	Mean	Std Error of Mean	Lower 95% CL for Mean	Upper 95% CL for Mean
d	100	-3.445794E-9	0.189105	-0.375225	0.375225

Alternatively, the IML procedure in SAS can be used to calculate the ratio estimate and its SE according to the formulas given. The following SAS program accomplishes the same task as the previous one. The program reads in the number of physicians and population for each county into vectors **y** and **x**, creates a new vector, **e**, defined as

$$\mathbf{e} = \mathbf{y} - \frac{\bar{y}}{\bar{x}}\mathbf{x}$$

gets the variance of **e**, and calculates the ratio estimate and SE according to the formula.

```
proc iml;
 use counties;
   /* define x and y */
 read all var{FARMPOP} into y;
 read all var{LANDAREA} into x;
 close counties;
```

```
 bign=3141;  /* designates the population size N (bign) */
 n=nrow(x);  /* get sample size  */
 xbar=sum(x)/n;
 ybar=sum(y)/n;
 bhat=ybar/xbar; /* the ratio estimate*/
 e=y-bhat*x;
 vard=(ssq(e)-(sum(e))**2/n)/(n-1);
 varbhat=((bign-n)/(bign*n))*vard/(xbar**2); /* estimated variance
                         for the estimate*/
 sebhat=sqrt(varbhat);   /* standard error */
 print " Ratio Estimation";
 print ybar xbar bhat varbhat sebhat;
quit;
run;
```

The output below gives a ratio estimate $\hat{B}$ of 1.2137 with the SE being 0.1891, which matches the results we obtained earlier using the first approach.

```
                    Ratio Estimation


        YBAR     XBAR       BHAT     VARBHAT     SEBHAT

     1146.87   944.92  1.2137218  0.0357606  0.1891048
```

Ratio estimation is also used to estimate the total and mean of a single variable to increase the precision of the estimates. Ratio estimation gives better performance than the regular estimation of the mean of y when y and the auxiliary variable x are linearly related, and specifically, when the data are well fit by a straight line through the origin. In the discussion of one-stage cluster sampling with an SRS of clusters later in the paper, an example is provided where the ratio estimate of the mean gives a smaller error variance than the unbiased estimate of the mean.

Regression Estimation

Although ratio estimation works best for data that are well fit by a straight line through the origin, regression estimation might be more suitable for data that scatter around a straight line with an intercept, which is modeled by $y = \beta_0 + \beta_1 x$.

With $\bar{x}_U$ assumed to be known, the regression estimator of $\bar{y}_U$ is determined by

$$\hat{\bar{y}}_{\text{reg}} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}_U \tag{10}$$

where $\hat{\beta}_0$ and $\hat{\beta}_1$ are the least squares regression coefficients.

The regression estimator is biased. Its standard error can be approximated as

$$\text{SE}\left(\hat{\bar{y}}_{\text{reg}}\right) = \sqrt{\left(1 - \frac{n}{N}\right)\frac{s_e^2}{n}} \tag{11}$$

where $e_i = y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)$.

One way to obtain the standard error is to calculate s_e^2 from the residual sum of squares in the regression analysis output. Another possibility is to, as in ratio estimation, create the variable e_i, and run PROC SURVEYMEANS on the new variable. Lastly, the estimation can be implemented through the IML environment. Using 'counties.dat', treat population as the auxiliary variable and estimate the total number of physicians in the United States, along with the standard error. The 1993 United States total population was estimated to be 255,077,536. The regression procedure in SAS (PROC REG) is used to obtain the ANOVA table and regression coefficients:

```
proc reg data=counties;
 model PHYSICIA=TOTPOP;
run;
```

The REG Procedure
Model: MODEL1
Dependent Variable: PHYSICIA

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	239521351	239521351	2068.16	<.0001
Error	98	11349761	115814		
Corrected Total	99	250871112			

```
Root MSE            340.31440  R-Square   0.9548
Dependent Mean      297.17000  Adj R-Sq   0.9543
Coeff Var           114.51842

                    Parameter Estimates

                    Parameter     Standard
Variable   DF        Estimate        Error    t Value  Pr > |t|

Intercept  1       -54.23128     34.89764      -1.55    0.1234
TOTPOP     1         0.00296   0.00006519      45.48    <.0001
```

with $\hat{t}_x = 255,077,536$, $\hat{\beta}_1 = 0.00296$, and $\hat{\beta}_0 = -54.23128$, the total number of physicians is estimated by

$$\hat{t}_{\text{yreg}} = \hat{t}_x\hat{\beta}_1 + N\hat{\beta}_0 = 255,077,536 \times 0.00296 \times 3141 \times (-54.23128) = 584689.0561$$

$$s_e = \sqrt{\frac{11349761}{n-1}} = \sqrt{\frac{11349761}{99}} = 338.5913 \ ,$$

so

$$\text{SE}\left(\hat{t}_{\text{yreg}}\right) = N\sqrt{1-\frac{n}{N}}\frac{s_e}{\sqrt{n}} = 3141\sqrt{1-\frac{100}{3141}}\frac{338.5913}{\sqrt{100}} = 104644.8691 \ .$$

The same result can be obtained by creating $e_i = y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)$ and applying the regular SURVEYMEANS procedure to obtain the sum of e_i and associated error. Suppose PROC REG was previously conducted, and coefficients $\hat{\beta}_0$ (-54.2313) and $\hat{\beta}_1$(0.0030) were obtained. The SAS code for obtaining the standard error of the regression estimator and its output are shown below:

```
data reg;
set counties;
wt=3141/100;
```

```
e=PHYSICIA-(-54.23128+TOTPOP*0.00296);
run;

proc surveymeans total=3141 sum;
var e;
weight wt;
run;
```

The SURVEYMEANS Procedure

Data Summary

Number of Observations	100
Sum of Weights	3141

Statistics

Variable	Sum	Std Dev
e	1724.554742	104648

Stratified Sampling with an SRS in Each Stratum

Stratified sampling means the population is divided into a number of mutually exclusive and exhaustive strata, and a probability sample is drawn from each stratum independently. The simplest form of stratified sampling where an SRS is taken from each stratum is of interest.

Let H denote the number of strata; N_h denote the number of sampling units in stratum h; $\bar{y}_h$ denote the sample mean in stratum h; and s_h^2 denote the sample variance in stratum h. The population total and mean are estimated by

$$\hat{t}_{\text{str}} = \sum_{h=1}^{H} N_h \bar{y}_h \ , \tag{12}$$

$$\bar{y}_{\text{str}} = \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_h \ . \tag{13}$$

Table 1. The number of plots sampled from each zone in seals.data

Zone	Number of Plots	Plots Sampled
1	68	17
2	84	12
3	48	11
Total	200	40

The variances of the estimates are

$$\hat{\mathrm{V}}\left(\hat{t}_{\mathrm{str}}\right)=\sum_{h=1}^{H}\left(1-\frac{n_h}{N_h}\right)N_h^2\,\frac{s_h^2}{n_h}\ , \tag{14}$$

$$\hat{\mathrm{V}}\left(\bar{y}_{\mathrm{str}}\right)=\sum_{h=1}^{H}\left(1-\frac{n_h}{N_h}\right)\left(\frac{N_h}{N}\right)^2\,\frac{s_h^2}{n_h}\ . \tag{15}$$

The approximate $100(1-\alpha)\%$ confidence interval for the mean is $\bar{y}_{\mathrm{str}} \pm z_{\alpha/2}\,\mathrm{SE}\left(\bar{y}_{\mathrm{str}}\right)$.

To implement stratified sampling analysis in PROC SURVEYMEANS, create a weight of N_h/n_h for each observation in stratum h. Other than the data file with the weight value in it, a file is needed that specifies stratum name and the corresponding N_h. Specifically, N_h should take on the variable name of '_total_' in this file. In PROC SURVEYMEANS, a statement specifying the name of the stratum variable should be added.

This example uses 'seals.dat' (Lohr, 1999, p. 123), which is on the number of breathing holes found in sampled areas of Svalbard fjords. The study was intended to estimate ringed seal populations. The study area was divided into three zones which define the strata. The total number of plots and the plots sampled in each zone are presented in Table 1.

The number of breathing holes in each sampled plot was recorded. Lohr (1999) showed the data as an example of post-stratification, where an SRS was taken from the entire population and then the number of units belonging to each stratum in the selected sample was recorded. For illustration purposes, these data are treated as a regular stratified sampling example because of its simplicity and accessibility. To estimate the total number of breathing holes in the study region, along with its standard error, we use the following SAS program:

```
data totals; /* This file gives Nh for each stratum */
input zone _total_ @@;
cards;
1 68 2 84 3 48
;
run;

data seals;
infile 'c:\sampling\seals.dat' dsd firstobs=2;
input zone holes;
if zone=1 then wt=68/17;
if zone=2 then wt=84/12;
if zone=3 then wt=48/11;
run;

proc surveymeans data=seals total=totals mean clm sum clsum;
    /* 'total=' specifies the name of the file containing Nh */
strata zone/list;
    /* the option 'list' gives more detailed information about
     each stratum */
var holes;
weight wt;
run;
```

This program leads to the following SAS output:

The SURVEYMEANS Procedure

Data Summary

Number of Strata	3
Number of Observations	40
Sum of Weights	200

Stratum Information

Index	Stratum zone	Population Total	Sampling Rate	N Obs	Variable	N
1	1	68	25.0%	17	holes	17
2	2	84	14.3%	12	holes	12
3	3	48	22.9%	11	holes	11

Statistics

Variable	Mean	Std Error of Mean	Lower 95% CL for Mean	Upper 95% CL for Mean	Sum	Std Dev
Holes	4.985909	0.590132	3.790188	6.181631	997.181818	118.026447

Variable	Lower 95% CL for Sum	Upper 95% CL for Sum
holes	758.037521	1236.326115

Because the total was being estimated, examine the sum. The number of breathing holes in the study region is estimated to be 997.1818 with standard error being 118.0264.

One-Stage Cluster Sampling with an SRS of Clusters

In cluster sampling, the population is divided into blocks, called clusters or primary sampling units (psus). Individual elements, which are secondary sampling units (ssus), are allowed in the sample only if they belong to a cluster that is included in the sample. Consider one-stage cluster sampling, where every element within a sampled cluster is included in the sample. Note that this just becomes an SRS with the units being the clusters and the variable on the unit being the total for the cluster.

The notation for cluster sampling is quite different than that for SRS and stratified sampling. It is defined as follows:

N = number of clusters or psus in the population
n = number of clusters or psus included in the sample
M_i = number of ssus in the i^{th} cluster
$K = \sum_{i=1}^{N} M_i$ = total number of ssus in the population
t_i = total in the ith cluster

There are two ways to estimate population totals and means: using unbiased estimation and using ratio estimation which is biased. Applying unbiased estimation,

$$\hat{t}_{\text{unb}} = \frac{N}{n}\sum_{i \in S} t_i \ , \tag{16}$$

$$\hat{\bar{y}}_{\text{unb}} = \frac{\hat{t}_{\text{unb}}}{K} \tag{17}$$

with variances estimated by

$$\hat{\text{V}}\left(\hat{t}_{\text{unb}}\right) = N^2\left(1-\frac{n}{N}\right)\frac{s_t^2}{n}\ , \tag{18}$$

where

$$s_t^2 = \frac{1}{n-1}\sum_{i\in S}\left(\hat{t}_i - \frac{\hat{t}_{\text{unb}}}{N}\right)^2\ ,$$

and

$$\hat{\text{V}}\left(\hat{\bar{y}}_{\text{unb}}\right) = \frac{1}{K^2}\hat{\text{V}}\left(\hat{t}_{\text{unb}}\right)\ . \tag{19}$$

When *K* is not known, only ratio estimation can be applied, which we will discuss later.

To obtain the unbiased estimate of the total in SAS, ignore the individual elements, and use the results for the sum in simple random sampling with the cluster totals as the observations and with weight being *N*/*n*. The unbiased estimate of population mean can be obtained in a similar way through scaling the cluster totals by 1/*K*.

The Green Globules data set (Lohr, 1999, p. 172) is used to demonstrate the SAS computation. The data set was originally a two-stage cluster sampling with SRS at each stage. Modifications to the data were made so that a one-stage cluster sampling scenario could be applied. Suppose the new candy Green Globules is being test marketed in an area of upstate New York. The market research firm decides to sample 6 of the 45 towns in the area and examine the number of cases of Green Globules sold in all supermarkets in the selected towns. The data set consists of two variables: town, which refers to the town the examined supermarket belongs to; and ncases, which refers to the number of cases sold in the examined supermarket. Suppose the total number of supermarkets is 252. The following SAS code reads in the data set and calculates unbiased estimates of the population total and mean.

```
data casesold;
 input town ncases @@;
 datalines;
 1 146  1 180  1 251  1 152  1 72  1 181  1 171  1 361
 1 73  1 186
 2 99  2 101  2 52  2 121
 3 199  3 179  3 98  3 63  3 126  3 87  3 62
 4 226  4 129  4 57  4 46  4 86  4 43  4 85  4 165
 5 12  5 23
 6 87  6 43  6 59
 ;
run;

proc means data=casesold nway;
 class town;
 var ncases;
 output out=tout sum=ts;
run;

data tvalue;
 set tout;
 newy=ts/252;
 wt=45/6;
run;

proc surveymeans data=tvalue total=45 sum;
 weight wt;
 var ts newy;
run;
```

As can be seen from the SAS code, the MEANS procedure was run to save the cluster totals into an output file "tout". The next step determines the weight, and defines the new variable "newy" by dividing the cluster totals by the total number of supermarkets, which is for the purpose of estimating the population mean. The SURVEYMEANS procedure requested the sum and produced the following output:

```
                    The SURVEYMEANS Procedure

                        Data Summary

                Number of Observations       6
                Sum of Weights              45
```

Statistics

Variable	Sum	Std Dev
ts	29565	10316
newy	117.321429	40.934698

The unbiased estimate of the population total (i.e., the total number of cases of Green Globules sold in the area) is 29565 with a SE of 10316. The estimate of the mean (i.e., the average number of cases of Green Globules sold in a supermarket in the area) is 117.3214 with a SE of 40.9347.

Using ratio estimation, the population mean and total are estimated by

$$\hat{\bar{y}} = \frac{\sum_{i \in S} t_i}{\sum_{i \in S} M_i} \ , \tag{20}$$

$$\hat{t}_r = K\hat{\bar{y}}_r \ , \tag{21}$$

with variances estimated by

$$\hat{\mathrm{V}}\left(\hat{\bar{y}}_r\right) = \left(1 - \frac{n}{N}\right) \frac{1}{n\bar{M}_U^2} \frac{\sum_{i \in S} M_i^2 \left(\bar{y}_i - \hat{\bar{y}}_r\right)^2}{n-1} \ , \tag{22}$$

$$\hat{\mathrm{V}}\left(\hat{t}_r\right) = N^2 \left(1 - \frac{n}{N}\right) \frac{1}{n} \frac{\sum_{i \in S} M_i^2 \left(\bar{y}_i - \hat{\bar{y}}_r\right)^2}{n-1} \tag{23}$$

In SAS, running PROC SURVEYMEANS directly with the cluster statement on the variable of interest produces the correct analysis for ratio estimation of the mean. Note however that choosing the sum option in the surveymeans procedure does not produce an estimate of the total.

```
proc surveymeans data=casesold total=45 mean;
 cluster town;
 var ncases;
run;
```

The SURVEYMEANS Procedure

Data Summary

Number of Clusters	6
Number of Observations	33

Statistics

Variable	Mean	Std Error of Mean
ncases	119.454545	21.384999

Note that the ratio estimate of the mean (i.e., 119.4545) takes a different value from the unbiased estimator, and in this example it has a smaller standard error of 21.3850.

Two-Stage Cluster Sampling with SRS at Each Stage

In two-stage cluster sampling, subsample only some of the elements of selected clusters. Consider an SRS of ssus is selected from each cluster. The same notations used with one-stage cluster sampling will be used, with the addition of m_i as the number of ssus chosen from the i^{th} cluster.

The individual psu total is estimated by

$$\hat{t}_i = \sum_{j \in S_i} \frac{M_i}{m_i} y_{ij} = M_i \bar{y}_i \tag{24}$$

Using unbiased estimation, the population total and mean are estimated by

$$\hat{t}_{\text{unb}} = \frac{N}{n} \sum_{i \in S} \hat{t}_i \ , \tag{25}$$

$$\hat{\bar{y}}_{\text{unb}} = \frac{\hat{t}_{\text{unb}}}{K} \ . \tag{26}$$

The variances of the estimates are estimated by

$$\hat{\mathrm{V}}\left(\hat{t}_{\mathrm{unb}}\right)=N^2\left(1-\frac{n}{N}\right)\frac{s_t^2}{n}+\frac{N}{n}\sum_{i\in S}\left(1-\frac{m_i}{M_i}\right)M_i^2\,\frac{s_i^2}{m_i}\ , \tag{27}$$

where

$$s_t^2=\frac{\sum_{i\in S}\left(\hat{t}_i-\frac{\hat{t}_{\mathrm{unb}}}{N}\right)^2}{n-1},\ s_i^2=\frac{\sum_{i\in S_i}\left(y_{ij}-\bar{y}_i\right)^2}{m_i-1}\ ,$$

and

$$\hat{\mathrm{V}}\left(\hat{\bar{y}}_{\mathrm{unb}}\right)=\frac{1}{K^2}\hat{\mathrm{V}}\left(\hat{t}_{\mathrm{unb}}\right)\ . \tag{28}$$

The unbiased estimate of the total can be obtained through the SURVEYMEANS procedure with the "cluster" option and weight of NM_i/nm_i. The variance can be seen as composed of two pieces, with the first piece being $N^2\left(1-\frac{n}{N}\right)\frac{s_i^2}{n}$ and the second piece being $\sum_{i\in S}\left(1-\frac{m_i}{M_i}\right)M_i^2\,\frac{s_i^2}{m_i}$:

$$\hat{\mathrm{V}}\left(\hat{t}_{\mathrm{unb}}\right)=\text{first piece}+\frac{N}{n}\left(\text{second piece}\right)$$

The first piece of the variance is the variance of the sum obtained through running the SURVEYMEANS procedure with the "cluster" statement and a weight of NM_i/nm_i. The second piece of the variance is given by the variance of the sum through running the SAS surveymeans procedure with the cluster variable specified in the "strata" statement and a weight of M_i/m_i. The example below demonstrates the details.

The example uses the data set named 'books.dat' (Lohr, 1999, p. 170). A home owner with a large library needs to estimate the purchase cost and replacement value of the book collection for insurance purposes. Twelve shelves were randomly selected from a total of 44 shelves, and 5 books were randomly selected from each of the selected shelves. This is a two-stage cluster sampling with

SRS at each stage. In this example, $N = 44$, $n = 12$, M_i is different for each bookshelf (M_is are given in Table 5.5 in the book), and $m_i = 5$. Suppose it is desired to estimate the total replacement value of the book collection. SAS code for manipulating the data set and conducting the analysis is as follows:

```
option ls=80 nodate;
data a;
 infile 'c:\sampling\books.dat' dlm=',' firstobs=2;
 input shelf number purchase replace;
 if shelf=2 then bigmi=26;
 if shelf=4 then bigmi=52;
 if shelf=11 then bigmi=70;
 if shelf=14 then bigmi=47;
 if shelf=20 then bigmi=5;
 if shelf=22 then bigmi=28;
 if shelf=23 then bigmi=27;
 if shelf=31 then bigmi=29;
 if shelf=37 then bigmi=21;
 if shelf=38 then bigmi=31;
 if shelf=40 then bigmi=14;
 if shelf=43 then bigmi=27;
 wt=44*bigmi/12/5;
 wt2=bigmi/5;
run;

/*to get the first piece of variance of t^ */
proc surveymeans data=a total=44 sum;
 cluster shelf;
 weight wt;
 var replace;
run;

/*to create a new file with _total_ being the cluster size Mi */
proc means data=a nway;
 class shelf;
 var bigmi;
 output out=ssize mean=_total_;
run;

/*to get the second piece of variance of t^ */
proc surveymeans data=a total=ssize sum;
 strata shelf/list;
 var replace;
 weight wt2;
run;
```

SAS output:

The SURVEYMEANS Procedure

Data Summary

Number of Clusters	12
Number of Observations	60
Sum of Weights	1382.33333

Statistics

Variable	Sum	Std Dev
replace	32638	5613.166224

The SURVEYMEANS Procedure

Data Summary

Number of Strata	12
Number of Observations	60
Sum of Weights	377

Statistics

Variable	Sum	Std Dev
replace	8901.200000	610.297665

The first piece of the variance is 5613.166224^2 and the second piece of the variance is 610.297665^2. Therefore,

$$\hat{t}_{\text{unb}} = 32637.73$$
$$\hat{\text{V}}\left(\hat{t}_{\text{unb}}\right) = 5613.166224^2 + \frac{44}{12}610.297665^2 = 32873333.6$$
$$\text{SE}\left(\hat{t}_{\text{unb}}\right) = \sqrt{\hat{\text{V}}\left(\hat{t}_{\text{unb}}\right)} = 5733.53$$

Again, ratio estimation could also be used with the population mean and total estimated by

$$\hat{\bar{y}}_r = \frac{\sum_{i\in S}\hat{t}_i}{\sum_{i\in S}M_i}\ , \tag{29}$$

$$\hat{t}_r = K\hat{\bar{y}}_r\ , \tag{30}$$

with

$$\hat{\text{V}}\left(\hat{\bar{y}}_r\right) = \frac{1}{\bar{M}^2}\left[\left(1-\frac{n}{N}\right)\frac{s_r^2}{n} + \frac{1}{nN}\sum_{i\in S}M_i^2\left(1-\frac{m_i}{M_i}\right)\frac{s_i^2}{m_i}\right]\ , \tag{31}$$

where

$$s_r^2 = \frac{\sum_{i\in S}\left(M_i\bar{y}_i - M_i\hat{\bar{y}}_r\right)^2}{n-1},\ s_i^2 = \frac{\sum_{i\in S_i}\left(y_{ij}-\bar{y}_i\right)^2}{m_i-1}\ ,$$

and

$$\hat{\text{V}}\left(\hat{t}_r\right) = K^2\,\hat{\text{V}}\left(\hat{\bar{y}}_r\right)\ . \tag{32}$$

As with the variance of the unbiased estimator, the variance of the ratio estimator can also be seen as the linear combination of two variance components. For the variance of the mean, for example, the first piece is $\frac{1}{\bar{M}^2}\left(1-\frac{n}{N}\right)\frac{s_r^2}{n}$ and the second piece is $\sum_{i\in S}M_i^2\left(1-\frac{m_i}{M_i}\right)\frac{s_i^2}{m_i}$, so that

$$\hat{\mathrm{V}}\left(\hat{\bar{y}}_r\right) = \text{first piece} + \frac{1}{\bar{M}^2 nN}\left(\text{second piece}\right) .$$

Note that the second piece takes exactly the same form as the second piece of the estimated variance of the unbiased total, which is given by the variance of the sum through running the SAS surveymeans procedure with the cluster variable specified in the strata statement and weight of M_i/m_i. The first piece can be seen as the sample variance of $\hat{e}_i = \frac{\hat{t}_i - M_i \hat{\bar{y}}_r}{\bar{M}}$. Therefore, the first piece can be obtained by creating this variable $\hat{e}_i$ and running the surveymeans procedure for SRS on the new variable in order to get the first component of the desired variance for the ratio estimator.

Use the previous 'books' for unbiased estimation in two-stage cluster sampling. To estimate the average replacement cost per book use ratio estimation:

```
/*Although the data are at ssu level, this will give Mbar as m is equal
across clusters*/
proc means mean data=a nway;
 var bigmi;
run;
```

```
Analysis Variable : bigmi

      Mean
31.4166667
```

The above SAS code gives the estimate of average cluster size $\bar{M} = 31.4167$. Using $\bar{M}$, the ratio estimate of the mean is

$$\hat{\bar{y}}_r = \frac{\hat{t}_{\text{unb}}}{\hat{K}} = \frac{32637.73}{31.416667 * 44} = 23.6106 .$$

The following SAS code and result lead to the variance of the ratio estimate of the mean:

```
/*The new variable gettihat is created here to facilitate getting ti's
in the next step */
data a;
 set a;
 gettihat=replace*bigmi/5;
```

```
run;

proc means data=a nway sum mean;
 class shelf;
 var gettihat bigmi;
 output out=ratio1 sum=tihat temp1 mean=temp2 bigmi;
run;

data ratio2;
 set ratio1;
 ei=(tihat-23.61061*bigmi)/31.4166667;
 keep shelf tihat bigmi ei;
run;

proc surveymeans data=ratio2 total=44 mean;
 var ei;
run;
```

```
                    The SURVEYMEANS Procedure

                          Data Summary

                  Number of Observations      12


                        Statistics

                                          Std Error
Variable                     Mean          of Mean
ei                  7.9575595E-8          5.410291
```

Therefore

$$\hat{\mathrm{V}}\left(\hat{\bar{y}}_r\right) = 5.410291^2 + \frac{1}{12*44*31.416667^2} 610.297665^2 = 29.9860$$

$$\mathrm{SE}\left(\hat{\bar{y}}\right) = 5.4760$$

Stratified Sampling with One-Stage Cluster Sampling (Using SRS) within Each Stratum

Consider stratified sampling with one-stage cluster sampling within each stratum. Let H denote the number of strata; N_h denote the total number of psus (i.e., clusters) within stratum h, and n_h denote the selected number of psus (or clusters) within

stratum h; t_{hi} denote the total for psu (cluster) i in stratum h, and t_h denote the total for stratum h; M_{hi} denote the number of individual observations (ssus) for cluster i within stratum h; s_{th}^2 denote the sample variance of t_{hi} in stratum h; K denote the total number of individual observations.

Horvitz-Thompson estimation

Using the Horvitz-Thompson procedure (Horvitz & Thompson, 1952), the population total and mean are estimated by:

$$\hat{t}_{HT} = \sum_h \frac{N_h}{n_h} \sum_{i \in S} t_{hi} \ , \tag{33}$$

$$\hat{\mathrm{V}}\left(\hat{t}_{HT}\right) = \sum_h N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{s_{th}^2}{n_h} \ . \tag{34}$$

When K is known,

$$\bar{y}_U = \frac{1}{K} \hat{t}_{HT} \ , \tag{35}$$

$$\hat{\mathrm{V}}\left(\hat{\bar{y}}_U\right) = \frac{1}{K^2} \hat{\mathrm{V}}\left(\hat{t}_{HT}\right) \ . \tag{36}$$

The implementation of cluster sampling under stratified sampling is straightforward in SAS. PROC SURVEYMEANS can be run with the stratum and cluster statements specified and the weight being N_h/n_h.

Table 2. Sampling design information for the ice cream spending data set

Grade	Number of Study Groups	Number of Students
7	608	1,824
8	252	1,025
9	403	1,151
Total	1,263	4,000

The data set used for illustration purpose comes from Example 96.1 of the SAS/STAT® 13.1 User's Guide (SAS Institute Inc., 2013). The study population is a junior high school with a total of 4,000 students in grades 7, 8, and 9. The variable of interest is how much these students spend weekly for ice cream. The clusters are study groups within grades. In each grade, a single stage cluster sampling is applied. Suppose the objective is to estimate t = the total amount spent by students (over all grades) and $\bar{y}_U$ = the mean spending per student, assuming that there are 4,000 total students. Table 2 shows the number of study groups and number of students in each grade.

Below is the SAS code that reads in the data set and conducts the analysis. Note that the variable StudyGroup identifies a student's study group. It is possible for students from different grades to have the same study group number because study groups are sequentially numbered within each grade.

```
option ls=80 nodate;
 data IceCreamStudy;
   input Grade StudyGroup Spending @@;
   datalines;
  7 34 7   7 34 7  7 412 4   9 27 14
  7 34 2   9 230 15  9 27 15   7 501 2
  9 230 8   9 230 7  7 501 3   8 59 20
  7 403 4   7 403 11  8 59 13   8 59 17
  8 143 12   8 143 16  8 59 18   9 235 9
  8 143 10   9 312 8  9 235 6   9 235 11
  9 312 10   7 321 6  8 156 19   8 156 14
  7 321 3   7 321 12  7 489 2   7 489 9
  7 78 1   7 78 10  7 489 2   7 156 1
  7 78 6   7 412 6  7 156 2   9 301 8
 ;
run;

data StudyGroups;
 input Grade _total_; datalines;
 7 608
 8 252
 9 403
 ;

data withweight;
 set IceCreamStudy;
 if grade=7 then wt=608/8;
 if grade=8 then wt=252/3;
 if grade=9 then wt=403/5;
run;
```

```
proc surveymeans sum data=withweight total=StudyGroups;
 strata Grade /list;
 cluster StudyGroup;
 var spending;
 weight wt;
run;
```

SAS output is presented below:

The SURVEYMEANS Procedure

Data Summary

Number of Strata	3
Number of Clusters	16
Number of Observations	40
Sum of Weights	3162.6

Stratum Information

Stratum Index	Grade	Population Total	Sampling Rate	N Obs	Variable	N	Clusters
1	7	608	1.32%	20	Spending	20	8
2	8	252	1.19%	9	Spending	9	3
3	9	403	1.24%	11	Spending	11	5

Statistics

Variable	Sum	Std Dev
Spending	28223	3456.556840

Therefore, $\hat{t}_{HT} = 28223$ and $\mathrm{SE}\left(\hat{t}_{HT}\right) = \sqrt{\hat{\mathrm{V}}\left(\hat{t}_{HT}\right)} = 3456.5568$. Further, $K = 4000$, $\bar{y}_U = \frac{1}{K}\hat{t}_{HT} = 7.0557$ and $\mathrm{SE}\left(\hat{\bar{y}}_U\right) = \frac{1}{K}\mathrm{SE}\left(\hat{t}_{HT}\right) = 0.8641$.

Ratio estimation As before, ratio estimation can also be carried out. The ratio estimator of the mean is determined by $\hat{\bar{y}}_r = \frac{\hat{t}}{\hat{K}}$. Although $\hat{t}$ is available from the demonstration of Horvitz-Thompson estimation in the early part of this section, $\hat{K}$ needs to be obtained to be able to compute the ratio estimate. $\hat{K}$ can be obtained in a similar way as $\hat{t}$ is obtained.

```
data withweight;
 set IceCreamStudy;
 size=1;
 if grade=7 then wt=608/8;
 if grade=8 then wt=252/3;
 if grade=9 then wt=403/5;
run;

data StudyGroups;
 input Grade _total_; datalines;
 7 608
 8 252
 9 403
 ;

proc surveymeans sum data=withweight total=StudyGroups;
 strata Grade /list;
 cluster StudyGroup;
 var size;
 weight wt;
run;
```

SAS output:

The SURVEYMEANS Procedure

Data Summary

Number of Strata	3
Number of Clusters	16
Number of Observations	40
Sum of Weights	3162.6

Stratum Information

Stratum Index	Grade	Population Total	Sampling Rate	N Obs	Variable	N	Clusters
1	7	608	1.32%	20	size	20	8
2	8	252	1.19%	9	size	9	3
3	9	403	1.24%	11	size	11	5

Statistics

Variable	Sum	Std Dev
size	3162.600000	237.486276

From the SAS output, $\hat{K} = 3162.6000$ [and $\mathrm{SE}\left(\hat{K}\right) = 237.4863$], which leads to the ratio estimate of $\hat{\bar{y}}_r = \frac{\hat{t}}{\hat{K}} = \frac{28223}{3162.6} = 8.9239$.

There are several approaches to obtain the variance. One approach defines the variance of $\hat{\bar{y}}_r$ as

$$\hat{\mathrm{V}}\left(\hat{\bar{y}}_r\right) = \sum_{h=1}^{H} N_h^2 \frac{N_h - n_h}{N_h n_h} s_e^2 \ , \tag{37}$$

where $e_i = \frac{t_{hi} - M_{hi}\hat{\bar{y}}_r}{\hat{K}}$.

The following SAS code creates the variable e_i and computes its sample variance, and hence the variance of $\hat{\bar{y}}_r$.

```
proc means sum n nway data=IceCreamStudy;
 class grade studygroup;
 var spending;
 output out=output1 sum=thi n=mhi;
run;

data ratio;
 set output1;
 if grade=7 then wt=608/8;
 if grade=8 then wt=252/3;
 if grade=9 then wt=403/5;
 ei=(thi-mhi*8.923860115)/3162.6;
run;

proc surveymeans sum varsum data=ratio total=StudyGroups;
 strata Grade /list;
 var ei;
 weight wt;
run;
```

SAS output:

The SURVEYMEANS Procedure

Data Summary

```
Number of Strata          3
Number of Observations    16
Sum of Weights          1263
```

Stratum Information

Stratum Index	Grade	Population Total	Sampling Rate	N Obs	Variable	N
1	7	608	1.32%	8	ei	8
2	8	252	1.19%	3	ei	3
3	9	403	1.24%	5	ei	5

Statistics

Variable	Sum	Std Dev	Var of Sum
ei	9.517492E-11	0.650859	0.423618

Therefore $\hat{\mathrm{V}}\left(\hat{\bar{y}}_r\right) = 0.4236$.

The second approach to obtain the variance of the ratio estimator makes use of Taylor's theorem (Woodruff, 1971) and specifies

$$\hat{\mathrm{V}}\left(\hat{\bar{y}}_r\right) = \frac{1}{\hat{K}^2}\left[\hat{\mathrm{V}}\left(\hat{t}\right) + \hat{\bar{y}}_U^2\,\hat{\mathrm{V}}\left(\hat{K}\right) - 2\hat{\bar{y}}_U\,\mathrm{Cov}\left(\hat{t},\hat{K}\right)\right] \tag{38}$$

$\hat{K}$, $\hat{\mathrm{V}}\left(\hat{t}\right)$, and $\hat{\mathrm{V}}\left(\hat{K}\right)$ were previously determined, and $\mathrm{Cov}\left(\hat{t},\hat{K}\right)$ must be estimated. This can be obtained through

$$\mathrm{Cov}\left(\hat{t},\hat{K}\right) = \sum_h N_h^2\left(1-\frac{n_h}{N_h}\right)\frac{\mathrm{Cov}\left(t_{hi},M_{hi}\right)}{n_h}\ . \tag{39}$$

Again using the Ice Cream data set, the following SAS code leads to $\mathrm{Cov}\left(\hat{t},\hat{K}\right)$:

```
/*the second approach to get the variance for ratio estimator*/
proc corr cov data=output1;
 var thi mhi;
 by grade;
run;
```

```
data covar1;
 input Grade covar nh bignh; datalines;
 7 2.42857143 8 608
 8 17.5 3 252
 9 6.45 5 403
 ;

data covar2;
 set covar1;
 cov=bignh**2*(1-nh/bignh)*covar/nh;
run;

proc means sum data=covar2 nway;
 var cov;
run;
```

SAS output:

Grade = 7
The CORR Procedure

2 Variables: thi mhi

Covariance Matrix, DF = 7

	thi	mhi
thi	37.71428571	2.42857143
mhi	2.42857143	0.28571429

Grade = 8
The CORR Procedure

2 Variables: thi mhi

Covariance Matrix, DF = 2

	thi	mhi
thi	358.3333333	17.5000000
mhi	17.5000000	1.0000000

Grade = 9

```
                    The CORR Procedure

                  2 Variables:  thi   mhi


                Covariance Matrix, DF = 4

                             thi           mhi

          thi    85.20000000    6.45000000
          mhi     6.45000000    0.70000000


                    The MEANS Procedure

                  Analysis Variable : cov

                            Sum
                         683681.12
```

From the previous SAS output,

$$\hat{\mathrm{V}}\left(\hat{t}\right)=11947785.19,\ \hat{K}=3162.6,\ \hat{\mathrm{V}}\left(\hat{K}\right)=56400$$

and

$$\mathrm{Cov}\left(\hat{t},\hat{K}\right)=\sum_h N_h^2\left(1-\frac{n_h}{N_h}\right)\frac{\mathrm{Cov}\left(t_{hi},M_{hi}\right)}{n_h}=683681.12\ .$$

Therefore

$$\begin{aligned}\hat{\mathrm{V}}\left(\hat{\bar{y}}_r\right)&=\frac{1}{\hat{K}^2}\left[\hat{\mathrm{V}}\left(\hat{t}\right)+\hat{\bar{y}}_u^2\,\hat{\mathrm{V}}\left(\hat{K}\right)-2\hat{\bar{y}}_u\,\mathrm{Cov}\left(\hat{t},\hat{K}\right)\right]\\&=\frac{1}{3162.6^2}\left[11947785.19+8.92386^2*56400-2*8.92386*683681.12\right]\\&=0.4236\end{aligned}$$

which is similar to what is obtained from the first approach.

One-Stage Cluster Sampling with Unequal Probabilities

Sometimes sampling using unequal probabilities can prove to be more efficient and provide more accurate estimates. Two new notations are added: π_i is the inclusion probability, the probability that the i^{th} cluster is in the sample; π_{ij} is the probability that clusters i and j are both in the sample. Because this is one-stage cluster sampling, whenever a cluster is selected, all units with a cluster are selected. The Horvitz-Thompson estimator of the population total is

$$\hat{t}_{HT} = \sum_{i \in S} \frac{t_i}{\pi_i} \ . \tag{40}$$

The variance can be estimated by

$$\hat{\mathrm{V}}\left(\hat{t}_{HT}\right) = \sum_{i \in S} \left(1-\pi_i\right) \frac{t_i^2}{\pi_i^2} + \sum_{i \in S} \sum_{\substack{k \in S \\ k \neq i}} \frac{\pi_{ik} - \pi_i \pi_k}{\pi_{ik}} \frac{t_i}{\pi_i} \frac{t_k}{\pi_k} \ . \tag{41}$$

SAS does not provide any built-in procedures to analyze sampling with unequal probabilities. The calculation of the variance of the Horvitz-Thompson estimator needs to be programmed using the IML procedure. On the other hand, the point estimate can be obtained directly by using the SURVEYMEANS procedure with the appropriate weight specified.

The analysis of sampling with probability proportional to size (pps) is illustrated for a one-stage sample using 'agpop.dat' (Lohr, 1999, p. 437). The data are from the U.S. 1992 Census of Agriculture. Only data from the state of Alabama are used for the purpose of illustration. The relevant variables are county, acres92 (i.e., the number of acres devoted to farms in 1992), and farms92 (i.e., the number of farms in 1992). The objective is to estimate the total acres in 1992 in the state of Alabama. Although data is available for all counties, first select 6 counties with probability proportional to size. Assume there is only acreage information on the 6 sampled counties, and the number of farms is assumed to be known previously for all counties. Below is the SAS code to select a pps sample and conduct the analysis based on the pps sample.

```
option ls=80 nodate;
data a;
infile 'C:\yingl\Sampling\agpop.dat' dlm=',' firstobs=2;
input COUNTY :$25. STATE $ ACRES92  ACRES87  ACRES82  FARMS92
```

```
FARMS87  FARMS82  LARGEF92  LARGEF87  LARGEF82  SMALLF92
SMALLF87  SMALLF82  REGION $;
if state='AL';
 /* convert missing values */
  array numval[*] _numeric_;
  do i = 1 to dim(numval);
  if numval[i]=-99 then numval[i]=.;
  end;
keep county state farms92 acres92;
run;

/* the sum of acres92 based on the original data set gives the true
population total */
proc means mean sum n;
var farms92 acres92;
run;
/* pps option in surveyselect procedure indicates sampling with
probability proportional to size, and jtprobs gives the joint
probabilities of selection */
proc surveyselect data=a out=cout method=pps sampsize=6 jtprobs;
size farms92;
run;
proc print data=cout;
run;
proc surveymeans data=cout mean sum;
var acres92;
weight SamplingWeight;
run;

proc iml;
use cout;
read all var{acres92} into t;
read all var{SelectionProb} into pi;
read all var{JtProb_1  JtProb_2  JtProb_3  JtProb_4  JtProb_5
JtProb_6}into jprob;
close cout;
n=nrow(t);
tht=t(t)*(1/pi);
var1=0;
var2=0;
do i =1 to n-1;
var1 = var1 + (t[i]**2)*(1 - pi[i])/(pi[i]**2);
do k =i+1 to n;
var2=var2+(2*t[i]*t[k]*(jprob[i,k]-
pi[i]*pi[k])/(pi[i]*pi[k]*jprob[i,k]));
end;
end;
var1 = var1 + (t[n]**2)*(1 - pi[n])/(pi[n]**2);
var=var1+var2;
```

```
se1=sqrt(var1);
se=sqrt(var);
print tht sel se var1 var2 var;
run;
```

SAS output:

The MEANS Procedure

Variable	Mean	Sum	N
FARMS92	565.7462687	37905.00	67
ACRES92	126131.69	8450823.00	67

The SURVEYSELECT Procedure

Selection Method PPS, Without Replacement
Size Measure FARMS92

Input Data Set	A
Random Number Seed	45721
Sample Size	6
Output Data Set	COUT

Obs	COUNTY	STATE	ACRES92	FARMS92	Selection Prob	Sampling Weight	Unit
1	LOWNDES COUNTY	AL	199714	315	0.04986	20.0556	1
2	WASHINGTON COUNTY	AL	85086	361	0.05714	17.5000	2
3	ETOWAH COUNTY	AL	85821	774	0.12252	8.1621	3
4	LIMESTONE COUNTY	AL	207226	910	0.14404	6.9423	4
5	BALDWIN COUNTY	AL	167832	941	0.14895	6.7136	5
6	LAUDERDALE COUNTY	AL	201892	1143	0.18093	5.5271	6

Obs	JtProb_1	JtProb_2	JtProb_3	JtProb_4	JtProb_5	JtProb_6
1	0.000000	0.002414	0.005175	0.006085	0.006292	0.007579
2	0.002414	0.000000	0.005940	0.006984	0.007222	0.008699
3	0.005175	0.005940	0.000000	0.015307	0.015829	0.019066
4	0.006085	0.006984	0.015307	0.000000	0.018849	0.022702
5	0.006292	0.007222	0.015829	0.018849	0.000000	0.023569
6	0.007579	0.008700	0.019066	0.022702	0.023569	0.000000

The SURVEYMEANS Procedure

Data Summary

Number of Observations	6
Sum of Weights	64.9007321

Statistics

Variable	Mean	Std Error of Mean	Sum	Std Dev
ACRES92	152173	28311	9876129	2914572

THT	SE1	SE	VAR1	VAR2	VAR
9876129.3	4651452.5	2934751.8	2.1636E13	-1.302E13	8.6128E12

Therefore

$$\hat{t}_{HT} = 9876129.3, \mathrm{SE}\left(\hat{t}_{HT}\right) = 2934751.8 \ .$$

General Complex Surveys

General complex surveys can usually be analyzed using Horvitz-Thompson estimation or ratio estimation given that the probability and joint probability of selection are obtainable. Given the extensive analyses needed for complex survey designs, SAS implementation is not given here. In general, when a survey involves several stages of cluster and stratified sampling, methodologies illustrated above can be applied stage-by-stage to conduct analyses from the bottom stage to the top stage.

For general illustration, consider a data set (Lohr, 1999) that relates to a 1991 nationwide survey conducted in the Gambia designed to estimate the prevalence of bed net use in rural areas.

> "The sampling frame consisted of all rural villages of fewer than 3000 people in The Gambia. The villages were stratified by three geographic regions (eastern, central, and western) and by whether the village had a public health clinic (PHC) or not. In each region five districts were chosen with probability proportional to the district population as estimated in the

> 1983 national census. In each district four villages were chosen, again with probability proportional to census population: two PHC villages and two non-PHC villages. Finally, six compounds were chosen more or less randomly from each village, and a researcher recorded the number of beds and nets, along with other information, for each compound." (p. 223)

The two variables of interest are Y = the number of beds with nets and X = the number of beds, and the sampling scheme used is

1. stratified sampling (region)
2. cluster sampling (district)
3. stratified sampling (PHC/non-PHC)
4. cluster sampling (village)
5. SRS (compound)

Let $\hat{t}y$ denote the estimated total for number of beds with nets, $\hat{t}x$ denote the estimated total for number of beds, M denote the total number of clusters, and m denote the number of clusters selected. Notations for subscripts are as follows: Let r denote region (r = 1, 2, 3), rd denote district d in region r, rdp denote PHC situation p (with 1 indicating PHC and 2 indicating non-PHC) in district d in region r, $rdpv$ denotes village v with PHC situation p in district d and in region r, and $rdpvc$ denotes compound c in village v with PHC situation p in district d and in region r.

Horvitz-Thompson estimation of the total Let π_i be the inclusion probability, the probability that the i^{th} unit is in the sample; and π_{ij} is the probability that units i and j are both in the sample. The Horvitz-Thompson procedure specifies that the calculation of the estimates or standard errors should start from the bottom stage up. During each stage of sampling,

$$\hat{t}_{HT} = \sum_{i \in S} \frac{\hat{t}_i}{\pi_i} \tag{42}$$

with variance of

$$\hat{\mathrm{V}}\left(\hat{t}_{HT}\right) = \sum_{i \in S} \left(1 - \pi_i\right) \frac{\hat{t}_i^2}{\pi_i^2} + \sum_{i \in S} \sum_{\substack{k \in S \\ k \neq i}} \frac{\pi_{ik} - \pi_i \pi_k}{\pi_{ik}} \frac{t_i}{\pi_i} \frac{t_k}{\pi_k} + \sum_{i \in S} \frac{\hat{\mathrm{V}}\left(\hat{t}_i\right)}{\pi_i} \ , \tag{43}$$

where $\hat{t}_i$ and $\hat{\mathrm{V}}\left(\hat{t}_i\right)$ are obtained from the sampling analysis at the lower stage.

The steps below give $\hat{t}y$ and $\hat{\mathrm{V}}\left(\hat{t}y\right)$, and the same procedure would be followed to get $\hat{t}x$ and $\hat{\mathrm{V}}\left(\hat{t}x\right)$. At the 5th stage, which is the lowest stage,

$$\hat{t}y_{rdpv} = \frac{M_{rdpv}}{6}\sum_{c=1}^{6} y_{rdpvc} \ , \tag{44}$$

$$\hat{\mathrm{V}}\left[\hat{t}y_{rdpv}\right] = M_{rdpv}^2 \frac{M_{rdpv}-6}{6M_{rdpv}} sy_{rdpvc}^2 \ . \tag{45}$$

At the 4th stage,

$$\hat{t}y_{rdp} = \sum_{v=1}^{2} \frac{\hat{t}y_{rdpv}}{\pi_v} \ , \tag{46}$$

where $\pi_v = \frac{POP_{rdpv}}{POP_{rdp}} \bullet 2$, which is the inclusion probability for village *v*, and which is proportional to census population.

$$\hat{\mathrm{V}}\left[\hat{t}y_{rdp}\right] = \sum_{v=1}^{2}\left(1-\pi_v\right)\frac{\hat{t}y_{rdpv}^2}{\pi_v^2} + \sum_{v=1}^{2}\sum_{k\neq v}\frac{\pi_{vk}-\pi_v\pi_k}{\pi_{vk}}\frac{\hat{t}y_{rdpv}\hat{t}y_{rdpk}}{\pi_v\pi_k} + \sum_{v=1}^{2}\frac{\hat{\mathrm{V}}\left[\hat{t}y_{rdpv}\right]}{\pi_v} \ , \tag{47}$$

where π_{vk} is the inclusion probability for both village *v* and village *k*.

At the 3rd stage:

$$\hat{t}y_{rd} = \sum_{p=1}^{2} \hat{t}y_{rdp} \ , \tag{48}$$

$$\hat{\mathrm{V}}\left[\hat{t}y_{rd}\right] = \sum_{p=1}^{2} \hat{\mathrm{V}}\left[\hat{t}y_{rdp}\right] \ . \tag{49}$$

At the 2nd stage:

$$\hat{ty}_r = \sum_{d=1}^{5} \frac{\hat{ty}_{rd}}{\pi_d} \ , \tag{50}$$

where $\pi_d = \frac{POP_{rd}}{POP_r} \bullet 5$, which is the inclusion probability for district *d*, and which is proportional to census population.

$$\hat{\mathrm{V}}\left[\hat{ty}_r\right] = \sum_{d=1}^{5} (1-\pi_d) \frac{\hat{ty}_{rd}^2}{\pi_d^2} + \sum_{d=1}^{5} \sum_{k \neq d} \frac{\pi_{dk} - \pi_d \pi_k}{\pi_{dk}} \frac{\hat{ty}_{rd} \hat{ty}_{rk}}{\pi_d \pi_k} + \sum_{d=1}^{5} \frac{\hat{\mathrm{V}}\left[\hat{ty}_{rd}\right]}{\pi_d} \ , \tag{51}$$

where π_{dk} is the inclusion probability for both village *v* and village *k*.

At the top stage:

$$\hat{ty} = \sum_{r=1}^{3} \hat{ty}_r \ , \tag{52}$$

$$\hat{\mathrm{V}}\left[\hat{ty}\right] = \sum_{r=1}^{3} \hat{\mathrm{V}}\left[\hat{ty}_r\right] . \tag{53}$$

Ratio estimation of the mean To estimate the population mean, ratio estimation can usually be used at the top/first level of the sampling design. The general formula for ratio estimation is as follows:

$$\hat{B} = \frac{\hat{ty}}{\hat{tx}} \ , \tag{54}$$

$$\hat{\mathrm{V}}\left(\hat{B}\right) = \frac{\hat{\mathrm{V}}\left(\hat{ty} - \hat{B}\hat{tx}\right)}{\hat{tx}^2} = \frac{1}{\hat{tx}^2} \sum_{r=1}^{3} \hat{\mathrm{V}}\left(\hat{ty}_r - \hat{B}\hat{tx}_r\right) . \tag{55}$$

To compute $\hat{\mathrm{V}}\left(\hat{B}\right)$, create a new variable $e_i = y_i - \hat{B}x_i$; then the estimation of $\hat{\mathrm{V}}\left(\hat{B}\right)$ reduces to the estimation of $\hat{\mathrm{V}}\left(\hat{B}\right) = \frac{1}{\hat{tx}^2} \hat{\mathrm{V}}\left[\hat{te}\right]$. The estimation of $\hat{\mathrm{V}}\left[\hat{te}\right]$ follows a similar procedure as the estimation of $\hat{\mathrm{V}}\left[\hat{ty}\right]$ as illustrated above.

Specifically, using the same example, obtain $\hat{\mathrm{V}}\left[\hat{te}\right]$ by conducting stage-by-stage analysis.

At the 5th stage, which is the lowest stage,

$$\hat{te}_{rdpv} = \frac{M_{rdpv}}{6}\sum_{c=1}^{6} e_{rdpvc} = \hat{ty}_{rdpv} - \hat{B}\hat{tx}_{rdpv} \ , \tag{56}$$

$$\hat{\mathrm{V}}\left[\hat{te}_{rdpv}\right] = M_{rdpv}^{2}\,\frac{M_{rdpv}-6}{6M_{rdpv}}\,se_{rdpvc}^{2} \ . \tag{57}$$

At the 4th stage:

$$\hat{te}_{rdp} = \sum_{v=1}^{2}\frac{\hat{te}_{rdpv}}{\pi_v} = \hat{ty}_{rdp} - \hat{B}\hat{tx}_{rdp} \ , \tag{58}$$

where $\pi_v = \frac{POP_{rdpv}}{POP_{rdp}}\bullet 2$, which is the inclusion probability for village *v*, and which is proportional to census population.

$$\hat{\mathrm{V}}\left[\hat{te}_{rdp}\right] = \sum_{v=1}^{2}(1-\pi_v)\frac{\hat{ty}_{rdpv}^{2}}{\pi_v^{2}} + \sum_{v=1}^{2}\sum_{k\neq v}\frac{\pi_{vk}-\pi_v\pi_k}{\pi_{vk}}\frac{\hat{te}_{rdpv}\hat{te}_{rdpk}}{\pi_v\pi_k} + \sum_{v=1}^{2}\frac{\hat{\mathrm{V}}\left[\hat{te}_{rdpv}\right]}{\pi_v} \ , \tag{59}$$

where π_{vk} is the inclusion probability for both village *v* and village *k*.

At the 3rd stage:

$$\hat{te}_{rd} = \sum_{p=1}^{2}\hat{te}_{rdp} = \hat{ty}_{rd} - \hat{B}\hat{tx}_{rd} \ , \tag{60}$$

$$\hat{\mathrm{V}}\left[\hat{te}_{rd}\right] = \sum_{p=1}^{2}\hat{\mathrm{V}}\left[\hat{te}_{rdp}\right] \ . \tag{61}$$

At the 2nd stage:

$$\hat{te}_r = \sum_{d=1}^{5} \frac{\hat{te}_{rd}}{\pi_d} = \hat{ty}_{rd} - \hat{B}\hat{tx}_{rd} \ , \tag{62}$$

where $\pi_d = \frac{POP_{rd}}{POP_r} \bullet 5$, which is the inclusion probability for district *d*, and which is proportional to census population.

$$\hat{\mathrm{V}}\left[\hat{te}_r\right] = \sum_{d=1}^{5}\left(1-\pi_d\right)\frac{\hat{te}_{rd}^2}{\pi_d^2} + \sum_{d=1}^{5}\sum_{k\neq d}\frac{\pi_{dk}-\pi_d\pi_k}{\pi_{dk}}\frac{\hat{te}_{rd}\hat{te}_{rk}}{\pi_d\pi_k} + \sum_{d=1}^{5}\frac{\hat{\mathrm{V}}\left[\hat{te}_{rd}\right]}{\pi_d} \ , \tag{63}$$

where π_{dk} is the inclusion probability for both village *v* and village *k*.

At the top stage:

$$\hat{te} = \sum_{r=1}^{3} \hat{te}_r \ , \tag{64}$$

$$\hat{\mathrm{V}}\left[\hat{te}\right] = \sum_{r=1}^{3} \hat{\mathrm{V}}\left[\hat{te}_r\right] \ . \tag{65}$$

Conclusion

Although complex sampling schemes could effectively reduce the time and financial resources required to estimate population characteristics in education, such sampling might also introduce error and bias if the data resulting from the sample designs were analyzed in an inappropriate way. With very few established software systems available to conduct analyses for complex sampling designs in a "point-and-click" way, it would be helpful to have reference documentation that gives instructions on how to use SAS to conduct sampling analyses by utilizing the current features associated with the SURVEYMEANS procedure as well as the SAS IML programming environment. This paper reviewed various sampling designs within the framework of probability sampling and provided documentation on how to use SAS to estimate means and proportions in different sampling designs.

Acknowledgements

The author is grateful to Professor John P. Buonaccorsi. His lectures on survey sampling helped shape the focus of this study. He also provided insightful comments and suggestions on an earlier version of this manuscript.

References

Horvitz, D. G., & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association, 47*(260), 663-685. doi: 10.1080/01621459.1952.10483446

Lohr, S. L. (1999). *Sampling: Design and analysis*. Pacific Grove, CA: Duxbury Press.

SAS Institute Inc. (2013). *SAS/STAT® 13.1 user's guide*. Cary, NC: SAS Institute.

Woodruff, R. S. (1971). A simple method for approximating the variance of a complicated estimate. *Journal of the American Statistical Association*, *66*(334), 411-414. doi: 10.1080/01621459.1971.10482279