SAS/SUDAAN Program Notes
Biostatistics: A Guide to Design, Analysis, and Discovery
Chapter 15: Analysis of Survey Data

Although some survey data analysis procedures are available in SAS 9.1, we need to use SUDAAN procedures for survey analyses illustrated in this chapter. We are assuming SAS-callable SUDAAN program is installed in your computer. SAS is used for data entry and management and SUDAAN analytic procedures are used to carry out the analysis.

Note 15.1 – Analysis of data in Example 15.4

The following program enters the data in Example 15.4: **prof** = the number of professional workers; **mph** = the number of workers with MPH degree; **wgt** = the sample weight, 7.5 (=60/8); **tot** = the total number of counties (60). A new variable, **totxmph**, is created by multiplying the number of professional workers in the population to **mph**.

The first analysis is to estimate the total with mph in the population using simple expansion procedure. DESCRIPT is a SUDAAN procedure for descriptive statistics, TOTALS is to estimate the population total, and DESIGN specifies sampling design (**wor** stands for sampling without replacement). TOTCNT specifies the total count of population size and this facilitate the finite population correction in variance estimation. NEST is used to specify the strata and primary sampling units and _ONE_ indicates that nesting is not used in this example. WEIGHT specifies the sample weight variable. VAR specifies the variable to be analyzed.

The second analysis is to estimate the total with mph in the population using ratio estimation procedure. RATIO is a SUDAAN procedure for ratio estimation. NUMER specifies the numerator variable in the ratio and DENOM specifies the denominator variable.

```
data ex154;
input prof mph wgt tot;
totxmph=1150*mph;
datalines;
21 14 7.5 60
18  8 7.5 60
 9  3 7.5 60
13  6 7.5 60
15  8 7.5 60
22 13 7.5 60
30 17 7.5 60
27 15 7.5 60
;
run;
proc descript totals design=wor;
totcnt tot;
nest _one_;
weight wgt;
var mph;
run;
proc ratio design=wor;
totcnt tot;
```

```
nest _one_;
weight wgt;
numer totxmph;
denom prof;
run;
```

The result of the first analysis is shown below. The estimated population total is 630 (Total) with standard error of 97.3 (SE Total):

```
Variance Estimation Method: Taylor Series (WOR)
by: Variable, One.
```

Variable		One
MPH	Sample Size	8
	Weighted Size	60.00
	Total	630.00
	SE Total	97.32
	Mean	10.50
	SE Mean	1.62
	Lower 95% Limit	
	Mean	6.66
	Upper 95% Limit	
	Mean	14.34

The result of the second analysis is shown below. The estimated population total is 623 (Ratio Est.) with the standard error of 29.9 (SE Ratio):

```
Variance Estimation Method: Taylor Series (WOR)
by: Variable, One.
```

Variable		One
TOTXMPH/PROF	Sample Size	8
	Weighted Size	60.00
	Ratio Est.	623.23
	SE Ratio	29.92
	Lower 95% Limit	
	Ratio	552.47
	Upper 95% Limit	
	Ratio	693.98

Note 15.2 – Suppopulation Analysis in Example 15.6

Two SAS/SUDAAN programs are shown below. The first program tries to perform a subpopulation analysis by selecting out African Americans from the data file. This program would not run because only one PSU remained in the 13th and 15th strata. The second program performs a subpopulation analysis using SUBPOPULATION command, which is a proper procedure.

Both program assumes the data file, **adult4,** on the web is copied on drive c. Data dictionary for this data file is also available on the web. This data file contains some missing values as noted in the dictionary. The analytical result mentioned in Example 15.6 used the same data file with the missing values imputed. Since the text does not deal with imputation issues, we decided not to use imputed values in this analysis. Therefore the analytic result shown below is slightly different from statistics shown in the text.

The first program attempts to calculate the average **bmi** for African Americans. SET command specifies the data file to be brought in. The first IF statement instructs to exclude missing values in **bmi** and the second IF statement selects only African Americans. Proc DESCRIPT is used to calculate descriptive statistics and DESIGN=**wr** specifies the sampling design to be sampling with replacement (under this design the finite population correction is not used). NEST specifies the nesting variables to be **stra** (stratum) and **psu** (primary sampling unit) and WEIGHT specifies the sampling weight variable. VAR specifies a variable to be analyzed.

```
libname ch15 'c:\';
data ex156;
set ch15.adult4;
if bmi<99;
if race=2;
proc descript design=wr;
nest stra psu;
weight wgt;
var bmi;
run;
```

As stated above, this program didn't run. The following error messages explain that variance can not be calculated because only one PSU is available in the 13th and 15th stratum.

```
DATA ERROR
 (Message 831)
There is a problem with nest variable STRA=13.000000 in record 2058
It has only one PSU whose value is 1.000000
(Message 830)
Cannot compute variance contribution for WR design when sample size is 1

DATA ERROR
 (Message 831)
There is a problem with nest variable STRA=15.000000 in record 2125
It has only one PSU whose value is 2.000000
(Message 830)
Cannot compute variance contribution for WR design when sample size is 1

Sample size is 1 error occurred 2 times
Sample size is 1 error occurred 2 times

SUDAAN processing halted.
```

The following is a proper program to perform to calculate the average **bmi** for African Americans. The program is similar to the program above. Instead of select out only African Americans, it uses the entire data file and accomplishes the subpopulation analysis by using

SUBPOPLN command. SUBPOPLN specifies the analysis to be restricted to **race**=2 (African Americans).

```
libname ch15 'c:\';
data ex156;
set ch15.adult4;
if bmi<99;
proc descript design=wr;
subpopln race=2;
nest stra psu;
weight wgt;
var bmi;
run;
```

The output from this program is shown below (slightly edited):

```
Number of observations read    :   9372    Weighted count :     9644
Observations in subpopulation  :   2847    Weighted count:     1075
Denominator degrees of freedom :     23

Variance Estimation Method: Taylor Series (WR)
For Subpopulation: RACE = 2
by: Variable, One.
-----------------------------------------------------
| Variable        |                   | One          |
-----------------------------------------------------
| BMI             | Sample Size       |         2847 |
|                 | Weighted Size     |      1075.27 |
|                 | Mean              |        27.28 |
|                 | SE Mean           |         0.18 |
|                 | Lower 95% Limit   |              |
|                 |  Mean             |        26.90 |
|                 | Upper 95% Limit   |              |
-----------------------------------------------------
```

Note 15.3 – Descriptive Analysis in Example 15.7

The following SAS/SUDAAN program calculates weighted means or proportions, standard errors, and design effects shown in Example 15.7. The structure of the program is similar to the program in Note 15.2. Again missing values are not used. Two new variables are created, **black** and **hispanic**, to calculate the percents of black and Hispanic populations. In order to get the design effects DEFF is entered before DESIGN in the PROC statement.

```
libname ch15 'c:\';
data ex157;
set ch15.adult4;
if educat<77;
if avsbp<777;
if bmi<99;
if vit<3;
if smoke<7;
black=0;
if race=2 then black=1;
hispanic=0;
```

```
if race=3 then hispanic=1;
if vit=2 then vit=0;
if smoke=2 then smoke=0;
proc discript deff design=wr;
nest stra psu;
weight wgt;
var age black hispanic educat avsbp bmi vit smoke;
run;
```

The slightly edited output from this program is shown below. The mean for **black**, **hispanic, vit**, and **smoke** should be interpreted as the proportion. Since the imputed values are not use, the numerical values in the output are slightly different from the text.

```
Number of observations read   :   9165    Weighted count:     9447
Denominator degrees of freedom:     23

Variance Estimation Method: Taylor Series (WR)
by: Variable, One.
```

Variable		One
AGE	Sample Size	9165
	Mean	43.42
	SE Mean	0.57
	Lower 95% Limit	
	Mean	42.25
	Upper 95% Limit	
	Mean	44.59
	DEFF Mean #4	9.46
BLACK	Sample Size	9165
	Mean	0.11
	SE Mean	0.01
	Lower 95% Limit	
	Mean	0.09
	Upper 95% Limit	
	Mean	0.13
	DEFF Mean #4	8.57
HISPANIC	Sample Size	9165
	Mean	0.05
	SE Mean	0.01
	Lower 95% Limit	
	Mean	0.04
	Upper 95% Limit	
	Mean	0.06
	DEFF Mean #4	9.13
EDUCAT	Sample Size	9165
	Mean	12.41
	SE Mean	0.12
	Lower 95% Limit	
	Mean	12.16
	Upper 95% Limit	
	Mean	12.65
	DEFF Mean #4	14.11

AVSBP	Sample Size	9165
	Mean	122.14
	SE Mean	0.37
	Lower 95% Limit	
	Mean	121.37
	Upper 95% Limit	
	Mean	122.91
	DEFF Mean #4	3.78
BMI	Sample Size	9165
	Mean	25.96
	SE Mean	0.12
	Lower 95% Limit	
	Mean	25.72
	Upper 95% Limit	
	Mean	26.20
	DEFF Mean #4	4.50
VIT	Sample Size	9165
	Mean	0.43
	SE Mean	0.01
	Lower 95% Limit	
	Mean	0.40
	Upper 95% Limit	
	Mean	0.46
	DEFF Mean #4	5.70
SMOKE	Sample Size	9165
	Mean	0.51
	SE Mean	0.01
	Lower 95% Limit	
	Mean	0.49
	Upper 95% Limit	
	Mean	0.54
	DEFF Mean #4	5.18

Note 15.4 - Contingency Table Analysis in Example 15.8

The following program calculates the proportion of vitamin use by three levels of education for the entire population and also for Hispanic Americans, shown in Example 15.8. Proc DESCRIPT is used along with TABLES statement. Note that TABLES is followed by SUBGROUP and LEVELS to specify the number of categories in **edu** (1=less than H.S.; 2=H.S. graduate; 3=some college)

```
libname ch15 'c:\';
data ex158;
set ch15.adult4;
if educat<77;
if vit<3;
edu=1;
if educat=12 then edu=2;
if educat>12 then edu=3;
if vit=2 then vit=0;
```

```
hispanic=0;
if race=3 then hispanic=1;
proc descript design=wr;
nest stra psu;
weight wgt;
var vit;
tables edu;
subgrups edu;
levels 3;
run;
proc descript design=wr;
suppopln hispanic=1;
nest stra psu;
weight wgt;
var vit;
tables edu;
subgrups edu;
levels 3;
run;
```

The slightly edited output of this program is shown below. Mean in the table is the proportion of vitamin use since the user is coded as 1 and nonuser as 0. The first table shows the analysis for the total population and the second table shows a subpopulation analysis for Hispanic Americans.

```
Variance Estimation Method: Taylor Series (WR)
by: Variable, EDU.
-------------------------------------------------------------------------------------------------
-
| Variable        |                 | EDU
|                 |                 | Total         | 1             | 2             | 3
|
-------------------------------------------------------------------------------------------------
-
|                 |                 |               |               |               |
|
| VIT             | Sample Size     |          9811 |          4065 |          3046 |          2700
|
|                 | Mean            |          0.43 |          0.33 |          0.40 |          0.51
|
|                 | SE Mean         |          0.01 |          0.02 |          0.02 |          0.02
|
|                 | Lower 95% Limit |               |               |               |
|
|                 |  Mean           |          0.40 |          0.30 |          0.36 |          0.47
|
|                 | Upper 95% Limit |               |               |               |
|
|                 |  Mean           |          0.45 |          0.37 |          0.44 |          0.56
|
-------------------------------------------------------------------------------------------------
-

For Subpopulation: HISPANIC = 1
by: Variable, EDU.
-------------------------------------------------------------------------------------------------
-
| Variable        |                 | EDU
|                 |                 | Total         | 1             | 2             | 3
|
```

VIT	Sample Size	2560	1584	571	405
	Mean	0.31	0.26	0.33	0.44
	SE Mean	0.02	0.02	0.02	0.04
	Lower 95% Limit				
	Mean	0.27	0.22	0.29	0.37
	Upper 95% Limit				
	Mean	0.35	0.31	0.37	0.51

Note 15.5 – Regression Analysis in Example 15.9

The following SAS/SUDAAN program performs a regression analysis shown in Example 15.9. The program assumes that the data file, **adult4**, on the web is copied onto drive c. SET command the data set to be brought in. The next four IF statements specify conditions for bringing the data. These conditions are to exclude observations with missing values in variables: **height**, **weight**, **avsbp**, and **vit**. The next two IF statements are to recode two variables: **vit** and **sex.** Vitamin users are coded as 1 and nonusers as 0. Males are coded as 1 (the text indicates males are coded as 0 but it is an error) and females as 0. Proc REGRESS instructs to perform a regression analysis assuming sample DESIGN = **wr** (sampling with replacement, this simply indicates that finite population correction is irrelevant). NEST specifies strata and primary sampling units used in the sampling. WEIGHT specifies sample weight variable. MODEL statement specifies a dependent variable (before = sign) and five independent variables.

```
libname ch15 'c:\';
data ex155;
set ch15.adult4;
if height<777;
if weight<777;
if avsbp<777;
if vit<3;
if vit=2 then vit=0;
if sex=2 then sex=0;
proc regress design=wr;
nest stra psu;
weight wgt;
model avsbp= height weight age sex vit;
run;
```

The analytic results are shown below. The highlighted results are summarized in Table 15.11.

```
Number of observations read       :   9235    Weighted count:     9488
Observations used in the analysis :   9235    Weighted count:     9488
Denominator degrees of freedom    :     23

Maximum number of estimable parameters for the model is  6
```

```
File  contains   46 Clusters
  46 clusters were used to fit the model
Maximum cluster size is 280 records
Minimum cluster size is  72 records

Weighted mean response is 122.159240

Multiple R-Square for the dependent variable AVSBP: 0.392471

Variance Estimation Method: Taylor Series (WR)
SE Method: Robust (Binder, 1983)
Working Correlations: Independent
Link Function: Identity
Response variable AVSBP: AVSBP
by: Independent Variables and Effects.
```

Independent Variables and Effects	Beta Coeff.	SE Beta	Lower 95% Limit Beta	Upper 95% Limit Beta	T-Test B=0	P-value T-Test B=0
Intercept	106.28	6.77	92.29	120.28	15.71	0.0000
HEIGHT	-0.40	0.10	-0.61	-0.19	-3.92	0.0007
WEIGHT	0.09	0.00	0.08	0.10	19.11	0.0000
AGE	0.60	0.01	0.57	0.63	45.58	0.0000
SEX	4.03	0.65	2.68	5.38	6.16	0.0000
VIT	-1.20	0.42	-2.06	-0.33	-2.85	0.0090

Contrast	Degrees of Freedom	Wald F	P-value Wald F
OVERALL MODEL	6	37759.46	0.0000
MODEL MINUS INTERCEPT	5	1134.62	0.0000
INTERCEPT	1	246.80	0.0000
HEIGHT	1	15.36	0.0007
WEIGHT	1	365.10	0.0000
AGE	1	2077.66	0.0000
SEX	1	37.88	0.0000
VIT	1	8.13	0.0090

Note 15.6 – Logistic Regression Analysis in Example 15.10

The following SAS/SUDAAN program performs a logistic analysis shown in Example 15.10. The structure of the program is similar to the program in Note 5. Note that three dummy variables are created, **male**, **edu1** and **edu2**. Proc RLOGIST is for used to call in SUDAAN logistic procedure.

```
libname ch15 'c:\';
data ex1510;
set ch15.adult4;
if educat<77;
if vit<7;
edu1=0;
if educat=12 then edu1=1;
edu2=0;
if educat>12 then edu2=1;
if vit=2 then vit=0;
```

```
male=0;
if sex=1 then male=1;
proc rlogist design=wr;
nest stra psu;
weight wgt;
model vit= male edu1 edu2;
run;
```

The output from this program is shown below. Statistics in the output is slightly different from the text because imputed values are not used.

```
Independence parameters have converged in 5 iterations

Number of observations read       :   9811    Weighted count:     9858
Observations used in the analysis :   9811    Weighted count:     9858
Denominator degrees of freedom    :     23

Maximum number of estimable parameters for the model is  4

Sample and Population Counts for Response Variable VIT
  0:  Sample Count     6037    Population Count      5629
  1:  Sample Count     3774    Population Count      4229

R-Square for dependent variable VIT (Cox & Snell, 1989): 0.036831

-2 * Normalized Log-Likelihood with Intercepts Only : 13402.11
-2 * Normalized Log-Likelihood Full Model           : 13033.94
Approximate Chi-Square (-2 * Log-L Ratio)           :   368.17
Degrees of Freedom                                  :        3
Note: The approximate Chi-Square is not adjusted for clustering.
      Refer to hypothesis test table for adjusted test.

Variance Estimation Method: Taylor Series (WR)
SE Method: Robust (Binder, 1983)
Working Correlations: Independent
Link Function: Logit
Response variable VIT: VIT
by: Independent Variables and Effects.
```

Independent Variables and Effects	Beta Coeff.	SE Beta	Lower 95% Limit Beta	Upper 95% Limit Beta	T-Test B=0	P-value T-Test B=0
Intercept	-0.46	0.08	-0.62	-0.29	-5.76	0.0000
MALE	-0.50	0.06	-0.62	-0.37	-8.40	0.0000
EDU1	0.25	0.09	0.07	0.44	2.89	0.0082
EDU2	0.76	0.09	0.58	0.95	8.42	0.0000

```
by: Contrast.
```

Contrast	Degrees of Freedom	Wald F	P-value Wald F
OVERALL MODEL	4	54.18	0.0000
MODEL MINUS INTERCEPT	3	69.36	0.0000
INTERCEPT	1	33.21	0.0000

MALE	1	70.57	0.0000
EDU1	1	8.38	0.0082
EDU2	1	70.95	0.0000

by: Independent Variables and Effects.

Independent Variables and Effects	Odds Ratio	Lower 95% Limit OR	Upper 95% Limit OR
Intercept	0.63	0.54	0.75
MALE	0.61	0.54	0.69
EDU1	1.29	1.08	1.55
EDU2	2.15	1.78	2.59