CROSSTAB Example #1

SUDAAN Statements and Results Illustrated

- SETENV optional statement
- CHISQ, LLCHISQ
- PRINT
- RFORMAT
- SEWGT option

Input Data Set(s): NHANES3S3.SAS7bdat

Example

Using data from the Third National Health and Nutrition Examination Survey (NHANES III), investigate the relationship of arthritis with gender, age, and race/ethnicity among adults. Test the null hypothesis that arthritis is not associated with each demographic variable.

Solution

The data set consists of adults aged 17 and older from ***NHANES III***, a cross-sectional sample survey of the civilian, non-institutionalized population aged 2 months or older, fielded during 1988-1994 All variables in this example are from the home interview component of NHANES III, and all six years of data are analyzed. Thus, the sample weight variable is WTPFQX6, and the stratification and PSU variables are SDPSTRA6 and SDPPSU6, respectively.

This example was run in SAS-Callable SUDAAN, and the SAS program and *.LST files are provided. Three two-way cross tabulations are requested on the TABLES statement (*i.e.*, one each of sex, age, and race/ethnicity with arthritis [yes/no]). Each of the variables on the TABLES statement appears on the CLASS statement (frequencies for each variable are displayed following the design summary in the results section). The SAS program is displayed in ***Exhibit 1***.

The TEST statement requests two tests of independence in R*C tables: CHISQ and LLCHISQ. Both are stratum-specific and test the null hypothesis of no association of arthritis with each demographic variable. Since no test statistics have been requested after the slash, the default Wald-*F* test statistic will be used to test both hypotheses.

The SETENV statement is optional: it manipulates the printout so that more columns are printed on a single page.

The PRINT statement is optional. It is used here to demonstrate how to manipulate and customize the printed results. Without the PRINT statement, a variety of default statistics would be produced, including everything requested here, with default labels and formats.

The PRINT statement is used in this example to request only the statistics of interest (row percentages and related statistics, default statistics from the stratum-specific hypothesis test STEST group), to change some of the default names for calculated quantities (e.g., NSUM changed to SAMSIZE), and to specify a variety of formats for printed statistics.

The RFORMAT statements associate the SAS formats with the variables used in the CROSSTAB procedure.

Exhibit 1. SAS-Callable SUDAAN Code

```
libname in v604 "c:\10winbetatest\examplemanual\crosstab";

options pagesize=70 linesize=80;
proc format;
  value yesno 1="1=Yes"
              2="2=No";
  value age 1="17-34"
            2="35-49"
            3="50-64"
            4="65-90+";
  value sex 1="1=Male"
            2="2=Female";
  value race 1="1=nH White"
             2="2=nH Black"
             3="3=Mex Amer"
             4="4=Other";

PROC CROSSTAB DATA=in.hanes3s3 FILETYPE=SAS DESIGN=WR;
NEST SDPSTRA6 SDPPSU6;
WEIGHT WTPFQX6;

CLASS AGEGRP4 HSSEX DMARETHN HAC1A;
TABLES (HSSEX AGEGRP4 DMARETHN)*HAC1A;
TEST CHISQ LLCHISQ;

SETENV ROWWIDTH=12 COLWIDTH=10 LABWIDTH=25;
PRINT  NSUM="SAMSIZE" WSUM="POPSIZE" ROWPER SEROW LOWROW UPROW /
       STEST=DEFAULT WSUMFMT=F9.0 SEROWFMT=F6.3 LOWROWFMT=F6.3 UPROWFMT=F6.3
       STESTVALFMT=F10.2;
rformat agegrp4 age.;
rformat hac1a yesno.;
rformat hssex sex.;
rformat dmarethn race.;

RTITLE "Estimate Prevalence of Arthritis, By Gender, Age, and Race/Ethnicity"
      "Test Null Hyp: Arthritis NOT Related to Age/Gender/Race-Eth";
RFOOTNOTE "NHANES-III, 1988-1994, JULY 1997 DATA RELEASE, ADULTS (17+)" ;
```

Exhibit 2. First Page of SUDAAN Output (SAS *.LST File)

```
                              S U D A A N
           Software for the Statistical Analysis of Correlated Data
          Copyright      Research Triangle Institute    December 2011
                                 Release 11.0


DESIGN SUMMARY: Variances will be computed using the Taylor Linearization
Method, Assuming a With Replacement (WR) Design
    Sample Weight: WTPFQX6
    Stratification Variables(s): SDPSTRA6
    Primary Sampling Unit: SDPPSU6


Number of observations read    :  20050    Weighted count :187647206
Denominator degrees of freedom :     49
```

Exhibit 2 shows that SUDAAN read in 20,050 adults from the dataset. The value of the sampling weight variable WTPFQX6 summed over these 20,050 adults is 187,647,206, an estimate of the average U.S. adult (aged 17+) civilian, non-institutionalized population during 1988-1994. The denominator degrees of freedom (ddf) for NHANES III is calculated by SUDAAN by its identification of 98 "pseudo-PSUs" and 49 "pseudo-strata" in the data set (i.e., 49 ddf = 98 PSUs – 49 strata).

Next, SUDAAN displays the frequencies of the CLASS variables (***Exhibit 3***).

Exhibit 3. CLASS Variable Frequencies (AGEGRP4)

```
Frequencies and Values for CLASS Variables

by: AGEGRP4.
-----------------------------------
AGEGRP4          Frequency     Value
-----------------------------------
Ordered
  Position:
  1                   6900     17-34
Ordered
  Position:
  2                   4496     35-49
Ordered
  Position:
  3                   3402     50-64
Ordered
  Position:
  4                   5252    65-90+
-----------------------------------
```

Exhibit 4. CLASS Variable Frequencies (HSSEX)

```
Frequencies and Values for CLASS Variables

by: Sex.
--------------------------------------
Sex              Frequency       Value
--------------------------------------
Ordered
  Position:
  1                   9401      1=Male
Ordered
  Position:
  2                  10649    2=Female
--------------------------------------
```

Exhibit 5. CLASS Variable Frequencies (DMARETHN)

```
Frequencies and Values for CLASS Variables

by: Race-ethnicity.
----------------------------------------
Race-
  ethnicity     Frequency         Value
----------------------------------------
Ordered
  Position:
  1                   8483    1=nH White
Ordered
  Position:
  2                   5486    2=nH Black
Ordered
  Position:
  3                   5306    3=Mex Amer
Ordered
  Position:
  4                    775       4=Other
----------------------------------------
```

Exhibit 6. CLASS Variable Frequencies (HAC1A)

```
Frequencies and Values for CLASS Variables

by: Doctor ever told you had: arthritis.
---------------------------------
Doctor ever
  told you
  had:
  arthritis     Frequency    Value
---------------------------------
Ordered
  Position:
  1                  4298    1=Yes
Ordered
  Position:
  2                 15748     2=No
---------------------------------
```

And then SUDAAN displays the results from the PRINT statement (***Exhibit 7***), below.

Exhibit 7. HSSEX*HAC1A Crosstabulation

```
Variance Estimation Method: Taylor Series (WR)

Estimate Prevalence of Arthritis, By Gender, Age, and Race/Ethnicity
Test Null Hyp: Arthritis NOT Related to Age/Gender/Race-Eth

by: Sex, Doctor ever told you had: arthritis.
```

Sex		Doctor ever told you had: arthritis		
		Total	1=Yes	2=No
Total	SAMSIZE	20046	4298	15748
	POPSIZE	187611487	32666641	154944847
	Row Percent	100.00	17.41	82.59
	SE Row Percent	0.000	0.510	0.510
	Lower 95% Limit			
	ROWPER	.	16.410	81.539
	Upper 95% Limit			
	ROWPER	.	18.461	83.590
1=Male	SAMSIZE	9399	1570	7829
	POPSIZE	89630819	11789474	77841345
	Row Percent	100.00	13.15	86.85
	SE Row Percent	0.000	0.640	0.640
	Lower 95% Limit			
	ROWPER	.	11.919	85.506
	Upper 95% Limit			
	ROWPER	.	14.494	88.081
2=Female	SAMSIZE	10647	2728	7919
	POPSIZE	97980668	20877167	77103501
	Row Percent	100.00	21.31	78.69
	SE Row Percent	0.000	0.591	0.591
	Lower 95% Limit			
	ROWPER	.	20.143	77.480
	Upper 95% Limit			
	ROWPER	.	22.520	79.857

```
NHANES-III, 1988-1994, JULY 1997 DATA RELEASE, ADULTS (17+)
```

The above table displayed in ***Exhibit 7*** contains 20,046 of the 20,050 adults in the data set; four adults who answered "don't know" to the arthritis question are excluded from the analysis. No subject has a missing value for gender, race/ethnicity, or age. Had the INCLUDE=MISSING option been included with the CLASS statement, the four records would have been retained in the CROSSTAB analysis as an additional column in the table. In the sample, 4,298 adults reported arthritis, and 15,748 did not. The estimated total number of adults in the population with arthritis is 32,666,641 (the sum of the value of WTPFQX6 for the 4298 sample adults), and the estimated total adult population is 187,611,487 (the sum of the value of WTPFQX6 for the 20,046 sample adults). The ratio of these two point estimates is the estimated percentage (prevalence) of adults with arthritis (*i.e.*, 32,666,641/187,611,487 = 17.41%). The estimated standard error for the point estimate 17.41% is 0.510%. A 95% confidence interval on the population prevalence of arthritis is (16.410%, 18.461%).

Females seem to have a higher prevalence of arthritis than do males—21% vs. 13%. The estimated standard error for estimated population totals can be requested in CROSSTAB by specifying SEWGT on the PRINT statement.

Exhibit 8. AGEGRP4*HAC1A Crosstabulation

Variance Estimation Method: Taylor Series (WR)

Estimate Prevalence of Arthritis, By Gender, Age, and Race/Ethnicity
Test Null Hyp: Arthritis NOT Related to Age/Gender/Race-Eth

by: AGEGRP4, Doctor ever told you had: arthritis.

AGEGRP4		Doctor ever told you had: arthritis		
		Total	1=Yes	2=No
Total	SAMSIZE	20046	4298	15748
	POPSIZE	187611487	32666641	154944847
	Row Percent	100.00	17.41	82.59
	SE Row Percent	0.000	0.510	0.510
	Lower 95% Limit			
	ROWPER	.	16.410	81.539
	Upper 95% Limit			
	ROWPER	.	18.461	83.590
17-34	SAMSIZE	6900	228	6672
	POPSIZE	71857480	2822848	69034632
	Row Percent	100.00	3.93	96.07
	SE Row Percent	0.000	0.409	0.409
	Lower 95% Limit			
	ROWPER	.	3.185	95.163
	Upper 95% Limit			
	ROWPER	.	4.837	96.815
35-49	SAMSIZE	4496	557	3939
	POPSIZE	53642570	6647246	46995324
	Row Percent	100.00	12.39	87.61
	SE Row Percent	0.000	0.650	0.650
	Lower 95% Limit			
	ROWPER	.	11.144	86.242
	Upper 95% Limit			
	ROWPER	.	13.758	88.856
50-64	SAMSIZE	3401	1072	2329
	POPSIZE	32114722	9555128	22559594
	Row Percent	100.00	29.75	70.25
	SE Row Percent	0.000	0.892	0.892
	Lower 95% Limit			
	ROWPER	.	27.993	68.424
	Upper 95% Limit			
	ROWPER	.	31.576	72.007
65-90+	SAMSIZE	5249	2441	2808
	POPSIZE	29996716	13641419	16355297
	Row Percent	100.00	45.48	54.52
	SE Row Percent	0.000	0.905	0.905
	Lower 95% Limit			
	ROWPER	.	43.664	52.699
	Upper 95% Limit			
	ROWPER	.	47.301	56.336

NHANES-III, 1988-1994, JULY 1997 DATA RELEASE, ADULTS (17+)

The TOTAL row for ***Exhibit 7*** is the same as the earlier table for gender (***Exhibit 8***). The estimated prevalence of arthritis increases with increasing age, from a low of 4% among adults aged 17-34 years, to a high of 45% among adults aged 65 years or older.

Exhibit 9. DMARETHN*HAC1A Crosstabulation

Variance Estimation Method: Taylor Series (WR)

Estimate Prevalence of Arthritis, By Gender, Age, and Race/Ethnicity
Test Null Hyp: Arthritis NOT Related to Age/Gender/Race-Eth

by: Race-ethnicity, Doctor ever told you had: arthritis.

Race-ethnicity		Doctor ever told you had: arthritis		
		Total	1=Yes	2=No
Total	SAMSIZE	20046	4298	15748
	POPSIZE	187611487	32666641	154944847
	Row Percent	100.00	17.41	82.59
	SE Row Percent	0.000	0.510	0.510
	Lower 95% Limit			
	ROWPER	.	16.410	81.539
	Upper 95% Limit			
	ROWPER	.	18.461	83.590
1=nH White	SAMSIZE	8480	2389	6091
	POPSIZE	142595429	26880246	115715183
	Row Percent	100.00	18.85	81.15
	SE Row Percent	0.000	0.691	0.691
	Lower 95% Limit			
	ROWPER	.	17.502	79.722
	Upper 95% Limit			
	ROWPER	.	20.278	82.498
2=nH Black	SAMSIZE	5485	1055	4430
	POPSIZE	20995070	3455547	17539523
	Row Percent	100.00	16.46	83.54
	SE Row Percent	0.000	0.718	0.718
	Lower 95% Limit			
	ROWPER	.	15.067	82.048
	Upper 95% Limit			
	ROWPER	.	17.952	84.933
3=Mex Amer	SAMSIZE	5306	745	4561
	POPSIZE	9827951	964747	8863204
	Row Percent	100.00	9.82	90.18
	SE Row Percent	0.000	0.490	0.490
	Lower 95% Limit			
	ROWPER	.	8.874	89.153
	Upper 95% Limit			
	ROWPER	.	10.847	91.126
4=Other	SAMSIZE	775	109	666
	POPSIZE	14193038	1366101	12826937
	Row Percent	100.00	9.63	90.37
	SE Row Percent	0.000	1.319	1.319
	Lower 95% Limit			
	ROWPER	.	7.281	87.379
	Upper 95% Limit			
	ROWPER	.	12.621	92.719

NHANES-III, 1988-1994, JULY 1997 DATA RELEASE, ADULTS (17+)

Again, the TOTAL row in ***Exhibit 9*** is the same as in the preceeding tables. The estimated prevalence of arthritis seems to vary by race/ethnicity, ranging from 10% for Mexican-Americans and the "other" group, to 16% and 19% for non-Hispanic blacks and non-Hispanic whites, respectively.

Exhibit 10. Stratum-Specific Hypothesis Tests (HSSEX*HAC1A)

```
Variance Estimation Method: Taylor Series (WR)

Estimate Prevalence of Arthritis, By Gender, Age, and Race/Ethnicity
Test Null Hyp: Arthritis NOT Related to Age/Gender/Race-Eth

  Test Statistics for Stratum-Specific Hypotheses
  Variable HSSEX by Variable HAC1A

by: Hypothesis Test, Test Statistic.
-----------------------------------------------------------------
Hypothesis Test
   Test Statistic                  DF   Test Value      P-Value
-----------------------------------------------------------------
CHISQ (Obs - Exp)
   Wald-F                           1       131.43       0.0000
LLCHISQ (Log-Lin Model)
   Wald-F                           1       115.80       0.0000
-----------------------------------------------------------------
NHANES-III, 1988-1994, JULY 1997 DATA RELEASE, ADULTS (17+)
```

The ***CHISQ*** test in ***Exhibit 10*** is analogous to the standard Pearson chi-square test for non-survey data (*i.e.*, a comparison of the "Observed" to the "Expected" number of persons per cell). "Expected" is calculated under the null hypothesis that arthritis and gender are statistically independent in the population. "Observed" and "Expected" refer to the estimated number of persons in the population, not the number of persons in the sample. Here, the null hypothesis is rejected, since the *p*-value is less than .0001. Thus, males and females differ significantly on the prevalence of arthritis. Based on the above two-way table of arthritis (HAC1A) and gender (HSSEX), females have a higher prevalence.

LLCHISQ tests the null hypothesis that the odds of arthritis in the population are the same for males and females. (*Odds* of arthritis is the probability of having arthritis divided by the probability of not having arthritis.) The null hypothesis is rejected. Based on the above two-way table of arthritis and gender, females have higher odds of arthritis.

The default Wald-*F* test statistic was used to test both hypotheses.

Exhibit 11. Stratum-Specific Hypothesis Tests (AGEGRP4*HAC1A)

```
Variance Estimation Method: Taylor Series (WR)

Estimate Prevalence of Arthritis, By Gender, Age, and Race/Ethnicity
Test Null Hyp: Arthritis NOT Related to Age/Gender/Race-Eth

  Test Statistics for Stratum-Specific Hypotheses
  Variable AGEGRP4 by Variable HAC1A

by: Hypothesis Test, Test Statistic.
-----------------------------------------------------------------
Hypothesis Test
   Test Statistic                  DF   Test Value      P-Value
-----------------------------------------------------------------
CHISQ (Obs - Exp)
   Wald-F                           3       300.16       0.0000
LLCHISQ (Log-Lin Model)
   Wald-F                           3       278.60       0.0000
-----------------------------------------------------------------
NHANES-III, 1988-1994, JULY 1997 DATA RELEASE, ADULTS (17+)
```

In ***Exhibit 11***, both tests reject the null hypothesis and indicate a significant relationship between age group (AGEGRP4) and arthritis (HAC1A). Neither test indicates which age groups differ significantly from each other, however. Further investigation of the CHISQ results could be done in DESCRIPT by using VAR and CATLEVEL for the arthritis variable and then using CONTRAST, PAIRWISE, or

DIFFVAR with the age group variable. Further investigation of the LLCHISQ result could be done in LOGISTIC or MULTILOG.

Exhibit 12. Stratum-Specific Hypothesis Tests (DMARETHN*HAC1A)

```
Variance Estimation Method: Taylor Series (WR)

Estimate Prevalence of Arthritis, By Gender, Age, and Race/Ethnicity
Test Null Hyp: Arthritis NOT Related to Age/Gender/Race-Eth

  Test Statistics for Stratum-Specific Hypotheses
  Variable DMARETHN by Variable HAC1A

by: Hypothesis Test, Test Statistic.
-----------------------------------------------------------------
Hypothesis Test
   Test Statistic                   DF   Test Value      P-Value
-----------------------------------------------------------------
CHISQ (Obs - Exp)
   Wald-F                            3        22.71       0.0000
LLCHISQ (Log-Lin Model)
   Wald-F                            3        36.63       0.0000
-----------------------------------------------------------------
NHANES-III, 1988-1994, JULY 1997 DATA RELEASE, ADULTS (17+)
```

Finally, both hypothesis tests reject the null hypothesis and indicate a significant relationship between race/ethnicity group (DMARETHN) and arthritis (HAC1A) (***Exhibit 12***). Neither test indicates which race/ethnicity groups differ significantly from each other. As mentioned in the above table for age group, a comparison of race/ethnic groups on prevalence of arthritis could be done in DESCRIPT or LOGISTIC or MULTILOG.