

VARGEN Example #3

SUDAAN Statements and Results Illustrated

- XCORR
- XMEAN
- PARAMETER
- SUBPOPX
- SUBGROUP

Input Data Set(s): DA32722P1.SAS7bdat

Example

Using NSDUH 2010 data, this example looks at the subpopulation that has tried both cigarettes and alcohol. The program computes the correlation between the age individuals first tried alcohol and the age cigarettes were first tried. It also computes the mean age for the first alcohol try and cigarette try, takes the difference of the two ages, and compares them by type of county the individuals reside in.

Solution

The data set is from the National Survey on Drug Use and Health (NSDUH) in 2010. The target population of NSDUH is the non-institutionalized civilian population of United States citizens aged twelve and older. In the public use data file there are 57,873 individual observations, which is the dataset used in this example.

The two variables of interest in this example are ALCTRY and CIGTRY. ALCTRY is the age that an individual first had a drink of alcohol, where a drink is classified as an entire drink excluding a sip or a taste. In the dataset used in this example, 70.19% of respondents gave a valid age for ALCTRY. 28.72% of respondents had never had a drink of alcohol (ALCTRY = 991) and the remaining 1.1% of individuals in the sample gave bad data (ALCTRY = 985), did not know (ALCTRY = 994), refused to answer (ALCTRY = 997), or left the question blank (ALCTRY = 998). All individuals with values of ALCTRY that are greater than 100 were excluded from the analysis.

CIGTRY is the age that an individual first tried a cigarette, be it all or part of one. In the dataset used in this example 51.73% of respondents gave a valid age for CIGTRY. 47.58% of respondents had never used cigarettes (CIGTRY = 991), and the remaining 0.69% of individuals in the sample gave bad data (CIGTRY = 985), did not know (CIGTRY = 994), or refused to answer (CIGTRY = 997). All individuals with values of CIGTRY greater than 100 were excluded from the analysis.

For this example we needed to restrict the analysis to only include individuals who gave a valid age for both ALCTRY and CIGTRY in order for the means of ALCTRY and CIGTRY to be more readily interpretable. A SUBPOPX statement was used to limit the analysis to observations with valid ages for the two variables, ALCTRY and CIGTRY both less than or equal to 100.

A SUBGROUP statement, followed by a LEVELS statement is used to separate the output by the type of county a respondent resides in with the variable COUTYP2. COUTYP2 is equal to 1 if the individual

resides in a large metro area, 2 if in a small metro area, and 3 if in a non-metro area. Since there are three values within COUTYP2, the LEVELS statement will have a value of 3 to include all levels within the COUTYP2 variable.

An XCORR statement is used to determine the population Pearson correlation between ALCTRY and CIGTRY for the subpopulation.

Two XMEAN statements are used in the program. The first computes the mean age that individuals in the subpopulation first had a drink of alcohol, ALCTRY. The second XMEAN statement computes the mean age that individuals first tried cigarettes, CIGTRY.

A PARAMETER statement utilizes the x-statistics created by the XMEAN statements to compute the difference between the average age individuals first tried alcohol and the average age they first tried cigarettes.

A PRINT statement is used to control what statistics are displayed in the output window. ESTIM shows the estimate of the statistic being created and SEESTIM is used to display the standard error of the estimate. LOWESTIM and UPESTIM are used to display the lower and upper bounds, respectively, of the 95% confidence interval for each parameter estimated.

The variable containing the sample weights is ANALWT_C. VESTR is the analysis stratum and VEREP is the analysis replicate. Both VESTR and VEREP are used in the NEST statement. The dataset needs to be sorted by the variables in the NEST statement prior to calling VARGEN.

This example was run in SAS-Callable SUDAAN. Note that the basic SUDAAN code is the same for both SAS-Callable and Standalone versions.

Exhibit 1. SAS-Callable SUDAAN Code (DESIGN=WR)

```
libname lib '\\rtints29\sudaan\Data\NSDUH 2010\';
filename in '\\rtints29\sudaan\Data\NSDUH 2010\32722-0001-Data.stc';

proc cimport infile=in library=lib;
run;

data NSDUH;
set lib.da32722p1;
options fmtsearch = (lib.formats);
run;

proc sort data=NSDUH;
by vestr verep;
run;

proc vargen data=NSDUH design = wr;
weight analwt_c;
nest vestr verep;
subpopx alctry <=100 and cigtry <= 100;
subgroup coutyp2;
levels 3;
xcorr alccorr : alctry , cigtry / name="Correlation between ALCTRY and CIGTRY";
xmean alcmean : alctry / name="Mean Age of First Drink of Alcohol";
xmean cigmean : cigtry / name="Mean Age of First Cigarette Try";
parameter alc_cig : alcmean - cigmean / name="Alcohol Age Mean Minus Cig Age Mean";
print estim seestim lowestim upestim;
run;
```

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      July 2012
                Release 11.0.0-testing-233

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

Number of observations read      : 57873      Weighted count :253619107
Observations in subpopulation  : 28073      Weighted count :152379079
Denominator degrees of freedom :      60
```

The results displayed in *Exhibit 2* indicate that there are 28,073 sample members in the subpopulation of individuals with valid values for both ALCTRY and CIGTRY as specified by the SUBPOPX statement in *Exhibit 1*. They represent 152,379,079 individuals in the target population.

Exhibit 3. XCORR of ALCTRY and CIGTRY Output

Variance Estimation Method: Taylor Series (WR)
 For VARGEN Testing: ALCTRY <= 100 AND CIGTRY <= 100
 Computing the Correlation between ALCTRY and CIGTRY, the Means, and Difference of Means
 Also Computes the Confidence Intervals for Each Statistics
 by: Variable, COUNTY TYPE.

| Variable COUNTY TYPE | Estimate | SE Estimate | Lower 95% Limit | Upper 95% Limit |
|--|----------|-------------|--------------------|--------------------|
| Correlation between ALCTRY and CIGTRY | | | | |
| Total | 0.39 | 0.02 | 0.36 | 0.43 |
| (1) Large Metro | 0.39 | 0.02 | 0.34 | 0.44 |
| (2) Small Metro | 0.42 | 0.04 | 0.35 | 0.49 |
| (3) Nonmetro | 0.36 | 0.03 | 0.30 | 0.43 |

Data from NSDUH 2010

The results in *Exhibit 3* show the output for the x-statistics generated by the XCORR statement on ALCTRY and CIGTRY and displays it by county type. As can be seen the estimate for the total subpopulation is a correlation of 0.39 with a standard error of 0.02. Thus there is a statistically significant positive correlation between the two variables suggesting that there is a positive relationship between the age someone first has a drink of alcohol and the age someone first tries cigarettes. By evaluating the confidence intervals, none of the county types have significantly different correlations from either each other or the subpopulation total.

Exhibit 4. XMEAN of ALCTRY Output

Variance Estimation Method: Taylor Series (WR)
 For VARGEN Testing: ALCTRY <= 100 AND CIGTRY <= 100
 Computing the Correlation between ALCTRY and CIGTRY, the Means, and Difference of Means
 Also Computes the Confidence Intervals for Each Statistics
 by: Variable, COUNTY TYPE.

| Variable COUNTY TYPE | Estimate | SE Estimate | Lower 95% Limit | Upper 95% Limit |
|---------------------------------------|----------|-------------|--------------------|--------------------|
| Mean Age of First Drink of Alcohol | | | | |
| Total | 16.52 | 0.05 | 16.43 | 16.62 |
| (1) Large Metro | 16.58 | 0.06 | 16.45 | 16.71 |
| (2) Small Metro | 16.43 | 0.07 | 16.29 | 16.56 |
| (3) Nonmetro | 16.52 | 0.10 | 16.32 | 16.71 |

Data from NSDUH 2010

The results in *Exhibit 4* show the output from the XMEAN statement on ALCTRY and display the results by county type. The estimate for the total subpopulation mean of ALCTRY is 16.52 years of age with a standard error of 0.05. All of the confidence intervals for the total and county type overlap, thus there are no significant differences between the three different county types.

Exhibit 5. XMEAN of CIGTRY Output

```

Variance Estimation Method: Taylor Series (WR)
For VARGEN Testing: ALCTRY <= 100 AND CIGTRY <= 100
Computing the Correlation between ALCTRY and CIGTRY, the Means, and Difference of Means
Also Computes the Confidence Intervals for Each Statistics
by: Variable, COUNTY TYPE.
-----
Variable
COUNTY TYPE          Estimate      SE Estimate      Lower 95%      Upper 95%
                        Limit          Limit
-----
Mean Age of First
Cigarette Try
Total                  15.81         0.05             15.70          15.92
(1) Large Metro       15.99         0.08             15.82          16.15
(2) Small Metro       15.65         0.08             15.48          15.81
(3) Nonmetro          15.52         0.11             15.31          15.73
-----
Data from NSDUH 2010

```

The results in *Exhibit 5* show the output from the XMEAN statement on CIGTRY and display the results by county type. The estimate for the total subpopulation mean is 15.81 with a standard error of 0.05. All of the county types' intervals overlap with the subpopulation total interval, thus there are no statistically significant differences from the total mean of CIGTRY amongst county type means. However, small metro and large metro county types do significantly differ from each other at the 5% level and non-metro types and large metro types do significantly differ at the 5% level as well. The individuals in large metro areas, who have also had a drink of alcohol at some time in their lives, appear to be trying cigarettes for the first time at a later age than people in small metro and non-metro areas.

Exhibit 6. PARAMETER Statement to Compute Difference between Means

Variance Estimation Method: Taylor Series (WR)
 For VARGEN Testing: ALCTRY <= 100 AND CIGTRY <= 100
 Computing the Correlation between ALCTRY and CIGTRY, the Means, and Difference of Means
 Also Computes the Confidence Intervals for Each Statistics
 by: Variable, COUNTY TYPE.

| Variable | Estimate | SE Estimate | Lower 95% Limit | Upper 95% Limit |
|--|----------|-------------|-----------------|-----------------|
| Alcohol Age Mean Minus Cig Age Mean | | | | |
| Total | 0.72 | 0.05 | 0.61 | 0.82 |
| (1) Large Metro | 0.59 | 0.08 | 0.43 | 0.76 |
| (2) Small Metro | 0.78 | 0.09 | 0.60 | 0.97 |
| (3) Nonmetro | 0.99 | 0.10 | 0.79 | 1.19 |

Data from NSDUH 2010

Exhibit 6 displays the results from the PARAMETER statement which computes the difference from the XMEAN statement on ALCTRY (*Exhibit 4*) and the XMEAN statement on CIGTRY (*Exhibit 5*). The estimate for the total subpopulation is 0.72 with a standard error of 0.05. As can be determined through the confidence interval generated and the estimate with its corresponding standard error, this is statistically significantly different from zero, which suggests that individuals in this subpopulation (those who had tried both alcohol and cigarettes) on average try cigarettes for the first time at a younger age than they try alcohol for the first time. This result holds true within all county types as can be seen by looking at both the estimates with their corresponding standard errors and the confidence intervals.

In *Exhibit 6* the only two county types which are statistically significantly different are large metro and non-metro, apart from those two types, all other confidence intervals overlap with each other. From these results it appears that, on average, for residents of non-metro counties the first dink of alcohol comes later relative to their first try of a cigarette than for residents of large metro counties.