Paper AA-10-2013

Analyzing Categorical Variables from Complex Survey Data Using PROC SURVEYFREQ

Taylor Lewis, University of Maryland, College Park, MD

ABSTRACT

This paper explores features available in PROC SURVEYFREQ to analyze categorical variables in a complex survey data set, where "complex" denotes a data set characterized by one or more of the following features: unequal weights, stratification, clustering, and finite population corrections. Using a real-world complex survey data set, this paper illustrates the necessary syntax to calculate descriptive statistics and conduct select bivariate analyses, such as tests of association and the computation of odds ratios and relative risk statistics. Given alongside the syntax examples is some discussion of the theoretical reasons certain "standard" statistical techniques like the chi-square test of association require modification(s) when applied to complex survey data.

INTRODUCTION

This paper is a follow-up to Lewis (2013), MWSUG 2013 paper number AA09 entitled "Analyzing Continuous Variables from Complex Survey Data Using PROC SURVEYMEANS." It is recommended that paper be read prior to or in combination with this one. Whereas PROC SURVEYMEANS is the complex survey companion procedure to PROC MEANS, PROC SURVEYFREQ is the companion to PROC FREQ. We begin the paper with a brief background section on the National Survey of Family Growth from which all examples are drawn. Following that is a brief discussion regarding how PROC SURVEYMEANS and PROC SURVEYFREQ produce analogous results for certain univariate, descriptive statistics. The only difference is one of scaling: PROC SURVEYMEANS reports in terms of proportions, while PROC SURVEYFREQ reports in terms of percentages. We then transition into bivariate analyses and complex survey data adaptations that have been proposed for certain popular tests of association such as the chi-square statistic. Following that is a section devoted to odds ratios and a class of "risk" statistics that can prove handy when data can be summarized in a 2 x 2 table. The concluding section illustrates multivariate capabilities of PROC SURVEYFREQ, albeit briefly. It is believed analysts will find modeling procedures such as PROC SURVEYREG, PROC SURVEYLOGISTIC, or PROC SURVEYPHREG more efficient for these types of analyses when complex survey data are at hand. Those procedures will not be discussed in this paper, however.

BRIEF BACKGROUND ON THE NATIONAL SURVEY OF FAMILY GROWTH

Examples in this paper are drawn from the responses of 12,279 women aged 15 – 44 who were interviewed as part of the 2006 – 2010 National Survey of Family Growth (NSFG). The NSFG (http://www.cdc.gov/nchs/nsfg.htm) is sponsored by the National Center for Health Statistics, a subagency of the Centers for Disease Control and Prevention. The objective of the NSFG is to collect data on marriage, divorce, pregnancy, contraceptive use, sexual behaviors, and attitudes about these topics from men and women aged 15 – 44. That is, the survey targets individuals in the U.S. of childbearing age. Historically, the survey has not been administered yearly, but periodically every 5 – 7 years. They term an administration a "cycle," of which there were six between 1973 and 2002. The NSFG went "continuous" in 2006, resulting in interviews of 10,403 men and 12,279 women, or a total of 22,682 individuals, that occurred between June 2006 and June 2010. The public-use data page for the survey contains links to download preformatted SAS® code that can be used to read in the raw data with minimal modifications: http://www.cdc.gov/nchs/nsfg/nsfg_2006_2010_puf.htm. All examples in this paper were derived from data posted to that page.

To limit interviewer travel costs associated with data collection, the NSFG sample design begins by randomly selecting of a set of geographically clustered units from the comprehensive list of these units that, taken together, cover the entire U.S. land area. Subsequent stages of sampling involve finer geographical units, households, and ultimately individuals. Note that despite the clusters' nested structure, we only need to concern ourselves with clustering and stratification occurring at the primary sampling stage. This information is identifiable by the distinct code combinations of the variables SEST and SECU. The "SE" stands for "sampling error"; SECU stands for "sampling error computation unit." There is also a variable WGTW1Q16 that can be used to ensure these 12,279 respondents properly reflect the entire population of women aged 15 – 44.

UNIVARIATE ANALYSES

DESCRIPTIVE STATISTICS

We begin this section with a simple example demonstrating the equivalence of PROC SURVEYFREQ and PROC SURVEYMEANS when estimating the totals and proportions of all *K* distinct values of a categorical variable. For simplicity, this is illustrated in the example below for the dichotomous indicator variable EVRMARRY for which a code of 1 indicates the respondent has been married at least once and 0 if the respondent has never married. Specifying a categorical variable in

the CLASS and VAR statements of PROC SURVEYMEANS is essentially equivalent to specifying the same variable in the TABLE statement of PROC SURVEYFREQ.

Both SURVEY procedures' output begins with a summary of the complex survey features. We can interpret the sum of weights as an estimate of the number of eligible population units. That is, the figure of 61.7 million corresponds to the number of females aged 15 – 44. This estimate and its standard error also appear in the "Total" row of the SURVEYFREQ output. Observe how the statistics labeled "Sum" and "Std Dev" in the SURVEYMEANS output match the statistics labeled "Weighted Frequency" and "Std Dev of Wgt Freq" in the SURVEYFREQ output. The same is true for "Mean" and "Std Err" and "Percent" and "Std Err of Percent," with the little wrinkle that the latter two have been multiplied by 100 and so the decimal point has shifted two positions to the left relative to the former.

```
proc format;
  value EVRMARRY
    0='NEVER MARRIED'
    1='MARRIED';
run;

proc surveymeans data=NSFG_0610_F nobs sum mean;
  strata SEST;
  cluster SECU;
  class EVRMARRY;
  var EVRMARRY;
format EVRMARRY EVRMARRY.;
weight WGTQ1Q16;
run;

proc surveyfreq data=NSFG_0610_F ;
  strata SEST;
  cluster SECU;
  table EVRMARRY;
format EVRMARRY EVRMARRY.;
weight WGTQ1Q16;
run;
```

The SURVEYFREQ Procedure

Data Summary

Number of Strata	56
Number of Clusters	152
Number of Observations	12279
Sum of Weights	61754741.1

Table of EVRMARRY

EVRMARRY	Frequency	Weighted Frequency	Std Dev of Wgt Freq	Percent	Std Err of Percent
NEVER MARRIED	6745	28850952	941508	46.7186	0.9175
MARRIED	5534	32903789	1294136	53.2814	0.9175
Total	12279	61754741	1936743	100.000	

The SURVEYMEANS Procedure

Data Summary

Number of Strata	56
Number of Clusters	152
Number of Observations	12279
Sum of Weights	61754741.1

Class Level Information

Class Variable	Levels	Values
EVRMARRY	2	NEVER MARRIED MARRIED

Statistics

Variable	Level	N	Mean	Std Error of Mean	Sum	Std Dev
EVRMARRY	NEVER MARRIED	6745	0.467186	0.009175	28850952	941508
	MARRIED	5534	0.532814	0.009175	32903789	1294136

We can see from the output that approximately 46.7%, or a total of about 28.8 million, women aged 15 – 44 have never been married. Because measures of variation match those produced by PROC SURVEYMEANS, it follows that the formulas outlined in Lewis (2013) carry forward to the present paper. For example, PROC SURVEYFREQ uses equations (1) and (2) defined in that paper to estimate the k^{th} category's total and variance. As this is true for virtually all descriptive statistics, to avoid being redundant, no specific formulas are given in the present subsection. The reader seeking more information can refer to that paper or the documentation.

The example above shows the statistics PROC SURVEYFREQ outputs by default. Additional statistical keywords are available after the slash in the TABLE statement, many of which are counterparts to those available in the PROC SURVEYMEANS statement. Table 1 summarizes the ones that are directly interchangeable.

PROC SURVEYFREQ TABLE Statement	**PROC SURVEYMEANS Statement**	**Output**
VAR	VAR	Variance of the estimated proportion/percentage
VARWT	VARSUM	Variance of the estimated total
CV	CV	Coefficient of variation of the estimated proportion/percentage
CVWT	CVSUM	Coefficient of variation of the estimated total
CL	CL	Confidence interval of the estimated proportion/percentage based on the significance level given in the ALPHA= option (default is ALPHA=.05) and complex survey degrees of freedom (# PSUs - # strata)
CLWT	CLSUM	Confidence interval of the estimated total based on the significance level given in the ALPHA= option (default is ALPHA=.05) and complex survey degrees of freedom (# PSUs - # strata)

Table 1. Summary of Comparable Statistical Keywords for Categorical Variables: PROC SURVEYMEANS Statement versus the TABLE statement of PROC SURVEYFREQ

The utter volume of statistics output by PROC SURVEYFREQ can be overwhelming at times. For instance, any of the keywords specified from Table 1 are output alongside all statistics output by default. You may wish to remain cognizant of a series of suppression options prefixed with NO (e.g., NOWT, NOFREQ), which can be used to reduce clutter in the listing. Table 2 summarizes some of the more useful ones we will utilize periodically in subsequent examples appearing in this paper. As with the statistical keywords, these are specified after the slash in the TABLE statement.

Option	Effect
NOFREQ	Suppresses unweighted counts
NOWT	Suppresses weighted counts
NOCELLPERCENT	Suppresses overall cell percentages
NOPERCENT	Suppresses all percentages
NOTOTAL	Suppresses row and column totals (when applicable)
NOSTD	Suppresses all standard errors of totals and percentages

Table 2. Summary of Output Suppression Options Available after the Slash in the TABLE statement of PROC SURVEYFREQ

To motivate use of some of the options displayed in Tables 1 and 2, suppose we were interested in estimating the distribution of the number of pregnancies experienced by the eligible women in the population. This is maintained in the variable PREGNUM, which ranges from 0 to 19. Note that this tally reflects all pregnancies that occurred, not strictly those culminating in a live birth (although the data can be subset accordingly if desired). Suppose further we were less interested in counts than we were percentages. The example below demonstrates how to restrict output to only percentages and 95% confidence limits. In effect, the NOFREQ and NOWT options after the slash in the TABLE statement suppress the estimated totals associated with the pregnancy number distribution. Without these options specified, the output tends to wrap onto multiple lines and make the analysis task more difficult.

```
proc surveyfreq data=NSFG_0610_F;
  strata SEST;
  cluster SECU;
  table PREGNUM / nowt nofreq cl;
weight WGTQ1Q16;
run;
```

Table of PREGNUM

PREGNUM	Percent	Std Err of Percent	95% Confidence Limits for Percent	
0	37.8801	1.0944	35.7076	40.0525
1	14.3533	0.4942	13.3724	15.3342
2	17.8230	0.6021	16.6277	19.0182
3	13.9525	0.5886	12.7841	15.1208
4	8.1313	0.3525	7.4316	8.8311
5	3.7789	0.2510	3.2806	4.2771
6	1.9862	0.2142	1.5610	2.4114
7	0.9768	0.1407	0.6975	1.2561
8	0.5992	0.1100	0.3809	0.8174
9	0.2445	0.0783	0.0891	0.4000
10	0.0478	0.0165	0.0150	0.0807
11	0.1046	0.0598	0.0000	0.2234
12	0.0192	0.0092	0.0009	0.0375
13	0.0618	0.0409	0.0000	0.1429
15	0.0034	0.0034	0.0000	0.0102
16	0.0075	0.0074	0.0000	0.0222
18	0.0298	0.0298	0.0000	0.0889
19	0.0002	0.0002	0.0000	0.0007
Total	100.000			

Examining the output we discover the vast majority of women have experienced four or fewer unique pregnancies. In fact, the prevalence of any distinct count after that point is less than 5%. Toward the tail end of the distribution these percentages become extremely small and the standard errors become almost as large as the estimate itself (i.e., the CV approaches 100%). Note how the lower end of several of the confidence intervals is truncated at 0, since negative values would be nonsensical. The problem of making proper inferences on extreme proportions has received considerable attention in the general statistical literature (Vollset, 1993; Brown et al., 2001). The crux of the issue is that the sampling distribution of proportions of rare characteristics can be far from normal. In lieu of the standard linear (or "Wald") confidence intervals, a number of alternatives have been proposed (Wilson, 1927; Clopper and Pearson, 1934) and adapted to complex surveys (Korn and Graubard, 1998; Korn and Graubard, 1999). A few of the more popular techniques were introduced in PROC SURVEYFREQ with the release of SAS Version 9.3. These are discussed in the next section.

ALTERNATIVE METHODS OF CONSTRUCTING CONFIDENCE INTERVALS FOR EXTREME PROPORTIONS DESCRIPTIVE STATISTICS

One appealing method is to make use a log-odds, or *logit*, transformation on the estimated proportion, whose sampling distribution tends to be closer to normal than the proportion itself. If we denote the estimated proportion $\hat{p}$, the logit transformation is $\log(\hat{p}/(1-\hat{p}))$, where log() represents the natural logarithm function.

The logit-transformed confidence interval is calculated in two steps. The first is to compute the endpoints on the logit scale as

$$CI(\text{logit}) = \left\{ \log\left(\frac{\hat{p}}{1-\hat{p}}\right) - \frac{t_{df,\alpha} \times se(\hat{p})}{\left(\hat{p}(1-\hat{p})\right)}, \log\left(\frac{\hat{p}}{1-\hat{p}}\right) + \frac{t_{df,\alpha} \times se(\hat{p})}{\left(\hat{p}(1-\hat{p})\right)} \right\} \quad (1)$$

where $t_{df,\alpha}$ is the $(1-\alpha/2)^{th}$ percentile of a t distribution with complex survey degrees of freedom df and significance level α, and $se(\hat{p})$ is the standard error of the proportion accounting for the complex survey design.

If we denote the two endpoints of the interval in equation 1 $logit_L$ and $logit_U$, the second step is to transform them back to proportions by

$$CI(\hat{p}) = \left\{ \exp\left(\frac{\text{logit}_L}{1+\text{logit}_L}\right), \exp\left(\frac{\text{logit}_U}{1+\text{logit}_U}\right) \right\} \quad (2)$$

Returning to the analysis in example above, we can have PROC SURVEYFREQ calculate the set of logit-transformed 95% confidence intervals by specifying CL(TYPE=LOGIT) after the slash in the TABLE statement. The PSMALL<=*p*> option available within the parentheses is also handy, as it will force the alternative method specified in the TYPE= option to be employed only when the estimated proportion is less p or greater than $(1 - p)$. For instance, the syntax below calls for linear confidence intervals unless the proportion is less than 0.1 or greater than 0.9, in which case logit-transformed confidence intervals are calculated. Asterisks and a corresponding footnote in the output reflect where the alternative method was employed. We can also observe how confidence intervals that are not necessarily symmetric about the point estimate and are far less likely to be truncated at 0.

```
proc surveyfreq data=NSFG_0610_F;
  strata SEST;
  cluster SECU;
  table PREGNUM / nowt nofreq cl(psmall=.1 type=logit);
weight WGTQ1Q16;
run;
```

Table of PREGNUM

PREGNUM	Percent	Std Err of Percent	95% Confidence Limits for Percent	
0	37.8801	1.0944	35.7076	40.0525
1	14.3533	0.4942	13.3724	15.3342
2	17.8230	0.6021	16.6277	19.0182
3	13.9525	0.5886	12.7841	15.1208
4	8.1313	0.3525	7.4585	8.8591*
5	3.7789	0.2510	3.3110	4.3099*
6	1.9862	0.2142	1.6027	2.4591*
7	0.9768	0.1407	0.7337	1.2995*
8	0.5992	0.1100	0.4161	0.8621*
9	0.2445	0.0783	0.1294	0.4615*
10	0.0478	0.0165	0.0241	0.0950*
11	0.1046	0.0598	0.0336	0.3251*
12	0.0192	0.0092	0.0074	0.0497*
13	0.0618	0.0409	0.0166	0.2297*
15	0.0034	0.0034	0.0005	0.0247*
16	0.0075	0.0074	0.0010	0.0536*
18	0.0298	0.0298	0.0041	0.2162*
19	0.0002	0.0002	0.0000	0.0017*
Total	100.000			

* Logit confidence limits are computed for percents outside the PSMALL range, 10% to 90%.

Note that specifying PSMALL (without providing a proportion threshold) invokes the default threshold of 0.25, and omitting the PSMALL option altogether forces the alternative method be applied to all confidence intervals.

The logit transformation is just one of several methods available in PROC SURVEYFREQ. Two other noteworthy adjustment techniques are TYPE=WILSON, a complex survey adjustment to the method of Wilson (1927) discussed in Korn and Graubard (1999), and TYPE=CP, an adaptation to complex survey data attributable to Korn and Graubard (1998) of the method originally proposed by Clopper and Pearson (1934). Because there is no evidence in the literature that any one method is superior (Rust and Hsu, 2007), we will not devote separate examples to these techniques. Consult the documentation for more details.

GOODNESS-OF-FIT TESTS

In addition to making inferences on the proportions or totals for any of the *K* distinct values of a categorical variable, one may occasionally wish to perform a joint hypothesis test on a vector of null proportions. To motivate an example of this, let us simplify the analysis of pregnancy numbers begun in the previous section by concerning ourselves with a collapsed version consisting of *K* = 5 categories: 0, 1, 2, 3, or 4 or more pregnancies. Examining the output from example above, perhaps we were interested in testing whether this set of observed proportions $\hat{\mathbf{p}} = \{\hat{p}_1 = 0.3788, \hat{p}_2 = 0.1435, \hat{p}_3 = 0.1782, \hat{p}_4 = 0.1395, \hat{p}_5 = 0.1600\}$ is significantly different from a hypothesized true population set of proportions $\mathbf{p}_0 = \{p_{01} = 0.4, p_{02} = 0.15, p_{03} = 0.15, p_{04} = 0.15, p_{05} = 0.15\}$. That is, we were interested in testing H_0: $\hat{\mathbf{p}} = \mathbf{p}_0$ vs. H_1: $\hat{\mathbf{p}} \neq \mathbf{p}_0$.

In general, under an assumed simple random sample RSRS) of size *n*, one can construct the following chi-square *goodness-of-fit* test statistic as

$$\chi^2_{SRS} = n \sum_{k=1}^{K} \frac{\left(\hat{p}_k - p_{0k}\right)^2}{p_{0k}} \qquad (3)$$

and assess significance by comparing it to $\chi^2_{df,\alpha}$, the $(1 - \alpha)^{th}$ percentile of a chi-square distribution with degrees of freedom $df = K - 1$ and desired α. If $\chi^2_{SRS} > \chi^2_{df,\alpha}$, one would reject the null hypothesis. An asymptotically equivalent test statistic is the *likelihood ratio* test statistic

$$G^2_{SRS} = 2n\sum_{k=1}^{K} \hat{p}_k \ln\left(\frac{\hat{p}_k}{p_{0k}}\right) \tag{4}$$

As with the chi-square statistic, the likelihood ratio test statistic can be referenced against a chi-square distribution with $df = K - 1$.

When data collected via a complex survey design, these tests are no longer abide by the same chi-square distribution under the null hypothesis. Rao and Scott (1981; 1984) proposed methods for adjusting them based on a generalized design effect (Kish, 1965) of sorts. The qualifier "generalized" is used because there are actually K distinct design effects, one for each proportion. Although the details of computing this factor, which we can denote *gdeff*, are somewhat complex, the remedy is straightforward to apply: we simply divide the standard chi-square goodness-of-fit test statistic by this factor as follows:

$$\chi^2_{R-S} = \frac{\chi^2_{SRS}}{gdeff} \tag{5}$$

or

$$G^2_{R-S} = \frac{G^2_{SRS}}{gdeff} \tag{6}$$

From here, the adjusted test statistics are assumed to be properly rescaled such that they can be referenced against a chi-square distribution with $df = K - 1$.

PROC SURVEYFREQ will compute the standard and adjusted statistics if we specify the CHISQ and LRCHISQ options after the slash in the TABLE statement as well as provide the null proportions using syntax TESTP=(*values*), where *values* are given as either proportions or percentages. (The TESTP= syntax is actually optional; without a formal declaration, SAS will assume the test is for equal proportions, or that $p_{0k} = 1/K$ for all K categories.) The example below demonstrates syntax to carry out our joint pregnancy number proportion example with null vector $\mathbf{p}_0 = \{p_{01} = 0.4, p_{02} = 0.15, p_{03} = 0.15, p_{04} = 0.15, p_{05} = 0.15\}$.

```
proc format;
  value PRG4PLUS
    0='NONE'
    1='1 PREGNANCY'
    2='2 PREGNANCIES'
    3='3 PREGNANCIES'
    other='4 OR MORE PREGNANCIES' ;
run;

proc surveyfreq data=NSFG_0610_F;
  strata SEST;
  cluster SECU;
  table PREGNUM / chisq lrchisq testp=(.4 .15 .15 .15 .15);
format PREGNUM PRG4PLUS.;
weight WGTQ1Q16;
run;
```

The SURVEYFREQ Procedure

Data Summary

Number of Strata	56
Number of Clusters	152
Number of Observations	12279
Sum of Weights	61754741.1

Table of PREGNUM

PREGNUM	Frequency	Weighted Frequency	Std Dev of Wgt Freq	Percent	Test Percent	Std Err of Percent
NONE	4741	23392728	1014910	37.8801	40.00	1.0944
1 PREGNANCY	1928	8863834	408854	14.3533	15.00	0.4942
2 PREGNANCIES	2095	11006517	492924	17.8230	15.00	0.6021
3 PREGNANCIES	1616	8616306	469753	13.9525	15.00	0.5886
4 OR MORE PREGNANCIES	1899	9875356	437134	15.9913	15.00	0.5399
Total	12279	61754741	1936743	100.000		

Rao-Scott Chi-Square Test

Pearson Chi-Square	99.4805
Design Correction	3.4389
Rao-Scott Chi-Square	28.9277
DF	4
Pr > ChiSq	<.0001
F Value	7.2319
Num DF	4
Den DF	384
Pr > F	<.0001

Sample Size = 12279

Rao-Scott Likelihood Ratio Test

Likelihood Ratio Chi-Square	96.0805
Design Correction	3.4389
Rao-Scott Chi-Square	27.9390
DF	4
Pr > ChiSq	<.0001
F Value	6.9848
Num DF	4
Den DF	384
Pr > F	<.0001

Sample Size = 12279

From the component of the output labeled “Rao-Scott Chi-Square Test,” we observe the unadjusted chi-square statistic is $\chi^2_{SRS} = 99.4805$, but with *gdeff* = 3.4389, the Rao-Scott adjusted test statistic is $\chi^2_{R-S} = 99.4805/3.4389 = 28.9277$. This is still highly significant, suggesting the null hypothesis should be rejected, although the adjusted value is notably smaller than what would be reported if we naively ignored the complex survey design. In practice, especially if the sample design involves clustering, we can expect *gdeff* to be larger than 1, so failing to make the adjustment could lead to erroneous, anti-conservative inferences. A similar line of reasoning applies for the likelihood ratio version of the goodness-of-fit test.

The only component of the output thus far not discussed is the “F Value” row. This is an *F*-distribution transformation of the Rao-Scott chi-square statistics Thomas and Rao (1987) and Lohr (1999) assert can exhibit improved stability under certain circumstances. The test statistic is formed by dividing Rao-Scott test statistic by *K* – 1 and referencing against an *F* distribution with *K* – 1 numerator degrees of freedom and denominator degrees of freedom equaling *K* – 1 times the complex survey degrees of freedom. So under the default Taylor series linearization approach, where the complex survey degrees of freedom equal the number of distinct PSUs minus the number of distinct strata, this explains why PROC SURVEYFREQ is referencing an *F* distribution with *K* – 1 = 4 numerator degrees of freedom and (152 – 56)*(5 – 1) = 384 denominator degrees of freedom.

The default adjustments that occur when the CHISQ and LRCHISQ options are specified are *first-order* corrections. Thomas and Rao (1987) also discuss *second-order* corrections, which go one step further by matching not only at the mean of the chi-square distribution, but also the variance. We will not go into the computational details here, but the gist is that the first-order adjusted test statistic is divided through once more by a second adjustment factor. PROC SURVEYFREQ will construct these second-order corrections if you specify CHISQ(SECONDORDER) or LRCHISQ(SECONDORDER). Lohr (1999) suggests these are warranted if the cell design effects vary appreciable. Although there are no guidelines for how much variability is too much variability, you can at least inspect the cell design effects in the output by specifying the DEFF option after the slash in TABLE statement.

BIVARIATE ANALYSES

The examples demonstrated through this point in the paper have been univariate in nature. In this section, we consider bivariate analyses in which two variables are specified in the TABLE statement and separated by an asterisk. Before launching into the details of any specific analysis, it is instructive to first see how the PROC SURVEYFREQ output is oriented. The default appearance is not the grid-like default appearance of PROC FREQ; rather, it more closely resembles PROC FREQ output when the CROSSLIST option is specified after the slash in the TABLE statement.

The example below is a simple bivariate analysis of current religious affiliation (RELIGION) and an indicator variable of whether the female respondent has ever been married (EVRMARRY). The raw codes for these two variables are assigned formats such that more meaningful labels appear in the output. Typically, the row dimension—specified first—is reserved for the explanatory variable and the column the outcome variable. The TABLE statement reads RELIGION*EVRMARRY in the present analysis, because it seems more appropriate to assess whether religious affiliation is useful in predicting the marriage indicator than the other way around.

```
proc format;
  value EVRMARRY
    0='NEVER MARRIED'
    1='MARRIED';

  value RELIGION
    1='NO RELIGION'
    2='CATHOLIC'
    3='PROTESTANT'
    4='OTHER RELIGIONS' ;
run;

proc surveyfreq data=NSFG_0610_F;
  strata SEST;
  cluster SECU;
  table RELIGION*EVRMARRY;
format RELIGION RELIGION.
       EVRMARRY EVRMARRY.;
weight WGTQ1Q16;
run;
```

Table of RELIGION by EVRMARRY

RELIGION	EVRMARRY	Frequency	Weighted Frequency	Std Dev of Wgt Freq	Percent	Std Err of Percent
NO RELIGION	NEVER MARRIED	1461	6085509	319257	9.8543	0.5051
	MARRIED	890	4997620	326499	8.0927	0.5079
	Total	2351	11083129	544368	17.9470	0.8409
CATHOLIC	NEVER MARRIED	1633	7171300	444581	11.6126	0.6515
	MARRIED	1502	8228192	522667	13.3240	0.7801
	Total	3135	15399492	879057	24.9365	1.2740
PROTESTANT	NEVER MARRIED	3166	13231597	547467	21.4260	0.8001
	MARRIED	2590	16261335	884708	26.3321	1.0946
	Total	5756	29492932	1215576	47.7582	1.4143
OTHER RELIGIONS	NEVER MARRIED	485	2362546	517054	3.8257	0.8008
	MARRIED	552	3416643	643641	5.5326	0.9781
	Total	1037	5779189	1128340	9.3583	1.7242
Total	NEVER MARRIED	6745	28850952	941508	46.7186	0.9175
	MARRIED	5534	32903789	1294136	53.2814	0.9175
	Total	12279	61754741	1936743	100.000	

The output consists of a tabular summarization of the $R = 4$ distinct categories of the row variable and $C = 2$ distinct categories of the column variable. The two dimensions combine for 4 x 2 = 8 distinct cells, but PROC SURVEYFREQ also outputs the row and column marginal statistics. For instance, the third row of output, the first "Total" row under the EVRMARRY column, provides marginal information on the first row. The last two rows provide marginal information on the two column categories. Aside from these rows, the table summarizes cell-specific frequencies and weighted totals as well as table-wide (overall) cell percentages.

Similarly to an argument made earlier, it would be redundant to provide formulas for the standard errors output for totals and proportions in bivariate analyses. PROC SURVEYFREQ uses the same principles laid out in the section on Totals and Means described in Lewis (2013). The only difference is that the indicator variable could be one defining a row, column, or particular cell. For instance, the overall percentage and standard error of Catholics who have never married could be ascertained by creating an indicator variable equaling 1 if the respondent was both Catholic and never married and 0 otherwise and specifying this variable in the VAR statement of PROC SURVEYMEANS. Of course, one would need to multiply the resulting statistics by 100 to convert them from the proportion scale back to the percentage scale.

The default output generated above does not foster the most straightforward evaluation of our original research question, whether religious affiliation is predictive of having ever been married. One way to informally answer gauge this would be to determine whether the column categories' percentages within specific rows percentages vary or are more or less constant and mirror the summary at the bottom of the table. Specifying the ROW option after the slash in the TABLE statement will provide these figures. The output is already quite inundated with statistics, however, so it might be prudent to also specify one or more of the suppression options summarized in Table 2, such as NOWT and/or NOFREQ.

The intent of the example above was merely to acquaint the reader with the default output orientation for bivariate analyses. In the next section, we will formalize the two-way table notation and explore some of the additional statistical tests available within PROC SURVEYFREQ to help answer questions such as the one posed here.

TESTS OF ASSOCIATION

The goodness-of-fit tests outlined above extend to two-way tables, although in this context they are more commonly referred to as *tests of association*. Instead of comparing the observed proportions to a user-defined null vector of proportions, these

tests are focused on determining whether the row factor and column factor are *independent*, which is to say the two factors are not associated with one another.

Suppose a row factor consisting of $r = 1,\ldots,R$ distinct categories is crossed with a column factor consisting of $c = 1,...,C$ distinct categories. Table 3 offers a visualization of how data might be tabulated under an SRS data collection design. The R x C cell counts are suffixed with the row and column category indicators (e.g., n_{12} is the count of cases defined by the first category of the row factor and second category of the column factor), and a dot symbolizes summing over the dimension placeholder (e.g., $n_{\bullet 2}$ represents the second column factor category summed over all rows).

		Column Factor				**Row Totals**
		1	2	…	C	
Row Factor	1	n_{11}	n_{12}	…	n_{1C}	$n_{1\bullet}$
	2	n_{21}	n_{22}	…	n_{2C}	$n_{2\bullet}$
	…	…	…	…	…	…
	R	n_{R1}	n_{R2}	…	n_{RC}	$n_{R\bullet}$
Column Totals		$n_{\bullet 1}$	$n_{\bullet 2}$	…	$n_{\bullet C}$	N

Table 3. Notation of a Bivariate Tabulation in an SRS Data Collection Design

Any of the R x C cells can be converted to proportions by dividing through by n, as can any of the $R + C$ marginal counts. The *Pearson chi-square test of association* proceeds by calculating R x C *expected* cell proportions as product of the row and column proportions. That is, the expected proportion for the r^{th} row and c^{th} column would be $p_{rc} = \left(\frac{n_{r\bullet}}{n}\right) x \left(\frac{n_{\bullet c}}{n}\right)$. If we denote the R x C observed cell proportions $\hat{p}_{rc} = \frac{n_{rc}}{n}$, the chi-square test of association is calculated as

$$\chi^2_{SRS} = n \sum_{r=1}^{R} \sum_{c=1}^{C} \frac{\left(\hat{p}_{rc} - p_{rc}\right)^2}{p_{rc}} \qquad (7)$$

It has the same structure general as the goodness-of-fit test, only the summation terms and associated subscripts differ slightly. The observed test statistic is referenced against $\chi^2_{df,\alpha}$, a chi-square random variate with $df = (R - 1) \times (C - 1)$ and significance level α. If $\chi^2_{SRS} > \chi^2_{df,\alpha}$ there is evidence the two factors are not independent. Similar reasoning applies for the *likelihood ratio test of association*, which is calculated as

$$G^2_{SRS} = 2n \sum_{r=1}^{R} \sum_{c=1}^{C} \hat{p}_{rc} \ln\left(\frac{\hat{p}_{rc}}{p_{rc}}\right) \qquad (8)$$

Table 4 shows the tabulation comparable to Table 3 when data is collected via a complex survey design. Instead of counts, cells and the margins are populated with weighted counts, which correspond to estimated totals for the entire population. This explains why they are symbolized by a capital letter and topped with a hat. The weighted counts are used to formulate the observed and expected proportions. For example, the expected proportion for the r^{th} row and c^{th} column under the null hypothesis of independence is $p_{rc} = \left(\frac{\hat{N}_{r\bullet}}{\hat{N}}\right) x \left(\frac{\hat{N}_{\bullet c}}{\hat{N}}\right)$, whereas the observed proportion is $\hat{p}_{rc} = \frac{\hat{N}_{rc}}{\hat{N}}$.

		Column Factor				Row Totals
		1	2	…	*C*	
Row Factor	1	$\hat{N}_{11}$	$\hat{N}_{12}$	…	$\hat{N}_{1C}$	$\hat{N}_{1\bullet}$
	2	$\hat{N}_{21}$	$\hat{N}_{22}$	…	$\hat{N}_{2C}$	$\hat{N}_{2\bullet}$
	…	…	…	…	…	…
	R	$\hat{N}_{11}$	$\hat{N}_{11}$	…	$\hat{N}_{RC}$	$\hat{N}_{R\bullet}$
Column Totals		$\hat{N}_{\bullet 1}$	$\hat{N}_{\bullet 2}$	…	$\hat{N}_{\bullet C}$	$\hat{N}$

Table 4. Notation of a Bivariate Tabulation in a Complex Survey Data Collection Design

Even with the expected and observed proportions using weighted figures from Table 4 in hand, the test statistics in equations 7 and 8 still require rescaling to account for the complex survey design. SAS uses a similar algorithm as discussed in the previous section regarding goodness-of-fit tests attributable to Rao and Scott (1981; 1984) and Thomas and Rao (1987). As in equations 5 and 6, the adjusted test statistics are formed by dividing the original test statistics by a generalized design effect.

To see how PROC SURVEYFREQ carries out these calculations, let us return to the analysis assessing whether the current religious affiliation of a woman respondent in the NSFG has any bearing on whether she has ever been married. The example below demonstrates syntax to perform the chi-square and likelihood ratio tests of association with a first-order design correction. It is structured similarly to what we saw previously. The two tests of association are requested with the CHISQ and LRCHISQ options after the slash in the TABLE statement. Although not demonstrated here, as before, second-order corrections can be requested by specifying CHISQ(SECONDORDER) or LRCHISQ(SECONDORDER)—see the documentation for more details.

```
proc format;
 value EVRMARRY
   0='NEVER MARRIED'
   1='MARRIED';

 value RELIGION
   1='NO RELIGION'
   2='CATHOLIC'
   3='PROTESTANT'
   4='OTHER RELIGIONS' ;
run;

proc surveyfreq data=NSFG_0610_F;
  strata SEST;
  cluster SECU;
  table RELIGION*EVRMARRY / chisq lrchisq;
format RELIGION RELIGION.
       EVRMARRY EVRMARRY.;
weight WGTQ1Q16;
run;
```

```
The SURVEYFREQ Procedure

Table of RELIGION by EVRMARRY

  Rao-Scott Chi-Square Test

Pearson Chi-Square    83.2411
Design Correction      2.3749

Rao-Scott Chi-Square  35.0500
DF                           3
Pr > ChiSq             <.0001

F Value               11.6833
Num DF                       3
Den DF                     288
Pr > F                 <.0001

     Sample Size = 12279


  Rao-Scott Likelihood Ratio Test

Likelihood Ratio Chi-Square  83.2142
Design Correction             2.3749

Rao-Scott Chi-Square         35.0387
DF                                 3
Pr > ChiSq                  <.0001

F Value                     11.6796
Num DF                            3
Den DF                          288
Pr > F                       <.0001

        Sample Size = 12279
```

We note the Rao-Scott chi-square statistic (35.05) is the Pearson chi-square statistic (83.24) divided by a generalized design effect (2.38). The likelihood ratio chi-square statistic is also adjusted accordingly. Both are referenced against a chi-square distribution with $df = (4 - 1) \times (2 - 1) = 3$. A rescaled test statistic based on an underlying F distribution is also given, paralleling what was already discussed. Even after adjusting for the complex survey design, all tests conclude there is an association between these two factors.

We close this section making brief mention of how there is also a *Wald chi-square test of association* available in PROC SURVEYFREQ, which can be output using the keyword WCHISQ. Heeringa et al. (2010, pp. 167 – 168) cite a few references suggesting the alternative tests of association presented above tend to perform better.

RISK STATISTICS AND ODDS RATIOS

For categorical data that can be displayed as a 2 x 2 table, PROC SURVEYFREQ offers a class of risk statistics (Section 2.2 of Agresti, 1996) that have an appealingly straightforward interpretation. The necessary orientation for these types of analyses is portrayed in Table 5 below. Once again, weighted totals are the basic ingredients for these statistics.

	Column Factor 1	Column Factor 2	**Row Totals**
Row Factor 1	$\hat{N}_{11}$	$\hat{N}_{12}$	$\hat{N}_{1\bullet}$
Row Factor 2	$\hat{N}_{21}$	$\hat{N}_{22}$	$\hat{N}_{2\bullet}$
Column Totals	$\hat{N}_{\bullet 1}$	$\hat{N}_{\bullet 2}$	$\hat{N}$

Table 5. Notation of a 2 x 2 Table Permitting Estimation of Risk Statistics and an Odds Ratio

The estimated probability of a case falling in the first column factor (c = 1) given it falls within the first row factor (r = 1) can be written as $\Pr(c = 1 | r = 1) = \hat{p}_{c=1|r=1} = \hat{N}_{11} / \hat{N}_{1\bullet}$. Borrowing terminology from the field of epidemiology, this is referred to as the estimated Column 1 *risk* for the first row. Alternatively, this can be interpreted as the estimated proportion of cases with the first column factor amongst all cases with the first row factor. Similarly, we can express the estimated Column 1 risk for the second row as $\Pr(c = 1 | r = 2) = \hat{p}_{c=1|r=2} = \hat{N}_{21} / \hat{N}_{2\bullet}$. If we differentiate some risk factor using the row dimension (e.g., smoker/non-smoker, treatment/placebo) and some target outcome using the first column, one estimator of interest would be the estimated risk difference defined as

$$\hat{r}^{1}_{diff} = \hat{p}_{c=1|r=1} - \hat{p}_{c=1|r=2} = \hat{N}_{11} / \hat{N}_{1\bullet} - \hat{N}_{21} / \hat{N}_{2\bullet} \quad (9)$$

PROC SURVEYFREQ will compute row-specific estimated risks and the estimator in equation 9 when the statistical keyword RISK is specified after the slash in the TABLE statement. It will also estimate a standard error and form a confidence interval around the difference. Significance of the two estimated risks' difference can be assessed by determining whether or not this interval encompasses 0. This is equivalent to testing whether the sample proportions of the two domains defined by the two row factors are significantly different from one another.

Another useful statistic is the estimated *relative risk*, which is defined as the ratio of the two estimated risks, or

$$\hat{r}^{1}_{rel} = \frac{\hat{p}_{c=1|r=1}}{\hat{p}_{c=1|r=2}} \quad (10)$$

A ratio of 1 implies the two estimated risks are equivalent, or that the estimated risk difference is 0. A ratio of 1.3 can be interpreted as meaning the first row estimated risk is 30% greater than the second row estimated risk. This statistic can be output by specifying RELRISK after the slash in the TABLE statement. Note that it is always computed as the first row over the second row, whereas the estimated risks and estimated risk difference are calculated for both columns when the RISK keyword is requested.

The next example illustrates how these statistics can facilitate an analysis of the relationship between ever being married (EVRMARRY) and ever using the birth control pill (PILLR) for NSFG-eligible females. Although it may sound a bit strange, we might partition and compare the estimated "risk" of ever using the birth control pill (cases where PILLR=1) based on whether or not one has been married. We can do this by defining EVRMARRY as the row factor and PILLR as the column factor. The ORDER=FORMATTED option in the PROC statement forces the row and column factors to be ordered according to their respective formats. (Assigning format labels that start with 1 or 2 is a way to control which appears first and which second.)

```
proc format;
value EVRMARRY
  1='1. MARRIED'
  0='2. NEVER MARRIED';
value PILLR
  1='1. YES'
  2='2. NO';
run;
proc surveyfreq data=NSFG_0610_F order=formatted;
  strata SEST;
  cluster SECU;
  table EVRMARRY*PILLR / nofreq nocellpercent row risk relrisk;
format EVRMARRY EVRMARRY.
       PILLR PILLR.;
```

```
weight WGTQ1Q16;
run;
```

The SURVEYFREQ Procedure

Table of EVRMARRY by PILLR

EVRMARRY	PILLR	Weighted Frequency	Std Dev of Wgt Freq	Row Percent	Std Err of Row Percent
1. MARRIED	1. YES	28475997	1180305	86.5432	0.8180
	2. NO	4427792	304082	13.4568	0.8180
	Total	32903789	1294136	100.000	
2. NEVER MARRIED	1. YES	16545294	632194	57.3475	1.3355
	2. NO	12305658	582853	42.6525	1.3355
	Total	28850952	941508	100.000	
Total	1. YES	45021292	1490396		
	2. NO	16733450	720287		
	Total	61754741	1936743		

Column 1 Risk Estimates

	Risk	Standard Error	95% Confidence Limits	
Row 1	0.8654	0.0082	0.8492	0.8817
Row 2	0.5735	0.0134	0.5470	0.6000
Total	0.7290	0.0079	0.7133	0.7448
Difference	0.2920	0.0158	0.2606	0.3233

Difference is (Row 1 - Row 2)

Sample Size = 12279

Column 2 Risk Estimates

	Risk	Standard Error	95% Confidence Limits	
Row 1	0.1346	0.0082	0.1183	0.1508
Row 2	0.4265	0.0134	0.4000	0.4530
Total	0.2710	0.0079	0.2552	0.2867
Difference	-0.2920	0.0158	-0.3233	-0.2606

Difference is (Row 1 - Row 2)

Sample Size = 12279

```
Odds Ratio and Relative Risks (Row1/Row2)

                                Estimate     95% Confidence Limits

Odds Ratio                        4.7832        4.0034      5.7149
Column 1 Relative Risk            1.5091        1.4353      1.5867
Column 2 Relative Risk            0.3155        0.2752      0.3617

                          Sample Size = 12279
```

Observe how the estimated risk statistics for the first and second row in the Column 1 Risk Estimates output component effectively match values reported under the "Row Percent" heading of the tabular summary of the raw data for lines where the formatted value of PILLR is "1. Yes." Approximately 86.5% of females who have ever married have used the birth control pill, whereas that figure is only 57.3% for females who have never married. The 95% confidence limits reported are the same endpoints of a confidence interval that would appear in the tabular summary if the CL option were specified after the slash in the TABLE statement. Similar reasoning translates to the Column 2 Risk Estimates output component and the PILLR line with formatted value "2. No" in the tabular summary.

The "Total" line in the risk estimates portion of the output is an estimate of the overall risk. For example, the overall risk for the first column is estimated as $\hat{N}_{\bullet 1} / \hat{N} =$ 45,021,292 / 61,754,741 = 0.729, which is to say the marginal percentage of females who have ever used birth control is 72.9%. The last line given is the estimated difference between the first and second row estimated risks (0.865 – 0.573 = 0.292). We find that the 95% confidence interval around this statistic does not contain zero, so the difference is significant. Females who have married at least once are significantly more likely than females who have never married to report ever using the birth control pill. The estimated relative risk is $\hat{r}^1_{rel} = \frac{0.865}{0.573} = 1.51$, which is provided toward the bottom of the output alongside an associated confidence interval. That is, females who have married at least once are 50% more likely to have taken the birth control pill at some point than females who have yet to marry.

Another measure of association output is the *odds ratio*, which is the ratio of the odds of falling in the first column given being in the first row to the comparable odds given being in the second row. There are numerous algebraic representations for how it can be estimated, including either

$$\hat{O} = \frac{\left(\hat{p}_{c=1|r=1}\right)/\left(1-\hat{p}_{c=1|r=1}\right)}{\left(\hat{p}_{c=1|r=2}\right)/\left(1-\hat{p}_{c=1|r=2}\right)} \tag{11}$$

or

$$\hat{O} = \frac{\hat{N}_{11}\hat{N}_{22}}{\hat{N}_{12}\hat{N}_{21}} \tag{12}$$

Since the odds themselves are always nonnegative, so is any ratio of them. A value of 1 indicates equivalence with respect to the odds, suggesting independence of the row and column factors. Values nearer 0 or much greater than 1 indicate a strong association. Note that the odds ratio is inverted if the order of the columns or rows is reversed. For instance, we note the estimated odds ratio in the output of example above is (28,475,997*12,305,658) / (4,427,792*16,545,294) = 4.78. If the two rows factors were reversed, the reported estimated odds ratio would become 1/(4.78) ≈ 0.21.

The odds ratio is often confused with and misinterpreted as the relative risk. A common misconception is to interpret an odds ratio of, say, 4.78 as meaning the probability of falling in the first column is 4.78 times greater for the first row cases relative to the second. That is actually the concept of the relative risk. As Agresti (1996, p. 25) notes, the two quantities are algebraically related to one another, but approximate equivalence only occurs when the risks in both rows are close to 0.

A formal way to use the odds ratio in tests of association is to form a confidence interval around the estimated odds ratio to see if it encompasses 1. PROC SURVEYFREQ outputs this interval by default. Note how the interval is asymmetric about the point estimate as it is with the relative risk—in contrast, the confidence intervals formed around the risk and risk difference point estimates are symmetric. Because the sampling distributions of the two ratios are often skewed, an interval is first

formed with respect to a natural logarithm transformation of the ratio then back-transformed. (So the interval is symmetric on the natural logarithm scale.) This is similar to the reasoning behind proposed alternatives for confidence intervals of rare proportions discussed above.

MULTI-WAY TABLES

PROC SURVEYFREQ can be utilized with three or more dimensions, but the result is a series of bivariate analyses, one for each distinct value of the first dimension(s) identified. For example, specifying TABLE VAR1*VAR2*VAR3 will result in a series of two-way tables of VAR2 crossed with VAR3, one for each category of VAR1. So VAR1 can be synonymously considered a control factor, page factor, or even a BY statement factor. If four variables were provided in the TABLE statement, the combination of distinct values derived from the first two variables serves this purpose, etc.

Suppose we sought to replicate the relative risk analysis in the example above, only now controlling for religiosity. This is accomplished with the syntax below. First, we create a format consisting of dichotomized categories of RELIGION. The second step is to add this variable as the leading dimension in the TABLE statement. As we see from the output, estimated relative risk statistics and odds ratios are reported for both 2 X 2 tables defined by the two formatted values of RELIGION.

```
proc format;
value RELIGION
  1='1. NOT RELIGIOUS'
  2,3,4='2. RELIGIOUS';
value EVRMARRY
  1='1. MARRIED'
  0='2. NEVER MARRIED';
value PILLR
  1='1. YES'
  2='2. NO';
run;

proc surveyfreq data=NSFG_0610_F order=formatted;
  strata SEST;
  cluster SECU;
  table RELIGION*EVRMARRY*PILLR / nofreq nocellpercent row relrisk;
format RELIGION RELIGION.
       EVRMARRY EVRMARRY.
       PILLR PILLR.;
weight WGTQ1Q16;
run;
```

Table of EVRMARRY by PILLR
Controlling for RELIGION=1. NOT RELIGIOUS

EVRMARRY	PILLR	Weighted Frequency	Std Dev of Wgt Freq	Row Percent	Std Err of Row Percent
1. MARRIED	1. YES	4394379	298137	87.9294	1.6778
	2. NO	603241	93070	12.0706	1.6778
	Total	4997620	326499	100.000	
2. NEVER MARRIED	1. YES	3965343	243061	65.1604	1.9271
	2. NO	2120166	158204	34.8396	1.9271
	Total	6085509	319257	100.000	
Total	1. YES	8359722	464881		
	2. NO	2723407	190781		
	Total	11083129	544368		

Odds Ratio and Relative Risks (Row1/Row2)

	Estimate	95% Confidence Limits	
Odds Ratio	3.8949	2.7673	5.4819
Column 1 Relative Risk	1.3494	1.2623	1.4426
Column 2 Relative Risk	0.3465	0.2600	0.4617

Sample Size = 12279

The SURVEYFREQ Procedure

Table of EVRMARRY by PILLR
Controlling for RELIGION=2. RELIGIOUS

EVRMARRY	PILLR	Weighted Frequency	Std Dev of Wgt Freq	Row Percent	Std Err of Row Percent
1. MARRIED	1. YES	24081618	1109402	86.2950	0.8687
	2. NO	3824551	271231	13.7050	0.8687
	Total	27906169	1210848	100.000	
2. NEVER MARRIED	1. YES	12579952	550313	55.2590	1.5192
	2. NO	10185492	531540	44.7410	1.5192
	Total	22765443	834085	100.000	
Total	1. YES	36661569	1370203		
	2. NO	14010043	655677		
	Total	50671612	1795977		

Odds Ratio and Relative Risks (Row1/Row2)

	Estimate	95% Confidence Limits	
Odds Ratio	5.0981	4.1980	6.1911
Column 1 Relative Risk	1.5616	1.4722	1.6565
Column 2 Relative Risk	0.3063	0.2649	0.3543

Sample Size = 12279

Whereas the overall estimated relative risk for the first column factor (ever using the birth control pill) was determined in prior example to be approximately 1.51, the figure is somewhat less for those who do not affiliate with any particular religion (1.35) and slightly higher for those who are religious (1.56). Examining the estimated risks from the “Row Percent” column of the

tabular summary portion of the output, we can attribute this to the finding that the risks for those who have never been married differ substantively based on religiosity. Specifically, those who are not affiliated with any particular religion have a risk of 0.652, whereas those who are religious have a risk of 0.553. The corresponding risks for those who have been married are much closer: 0.879 versus 0.863.

For brevity, the example above pertained strictly to an analysis of relative risks, but the same concepts translate to other bivariate analyses. For example, resubmitting the PROC SURVEYFREQ step with the CHISQ option also requested after the slash in the TABLE statement will result in two Rao-Scott-adjusted chi-square test statistics, one for each of the 2 x 2 tables.

CONCLUSION

This paper introduced the capabilities of PROC SURVEYFREQ for a variety of categorical variable analyses. For basic descriptive statistics, we observed how PROC SURVEYFREQ and PROC SURVEYMEANS produce equivalent results, albeit sometimes on a different scale (i.e., percentages versus proportions). Because of this, no specific measures of variability were defined; they are the same as those defined in Lewis (2013). It was stressed that goodness-of-fit tests and tests of association discussed in general statistics texts such as Agresti (1996) can be used with weighted proportions generated from complex survey designs, but only after one or more adjustments factors illustrated by Rao and Scott (1981; 1984) are applied. These adjustments are available within PROC SURVEYFREQ, and are reported by default in the output.

The paper concluded with a brief section on multi-way tables, those consisting of three or more dimensions. Although these types of analyses can be useful for detecting intricate trends, a more common approach for multivariate analyses such as this is to fit and interpret a regression model. Depending on the nature and scale of the outcome variable, three customized procedures available for this purpose when complex survey data are at hand are PROC SURVEYREG, PROC SURVEYLOGISTIC, and PROC SURVEYPHREG. These procedures were not demonstrated in this paper, but readers seeking more information are referred to Lewis (2012) and Berglund (2011).

REFERENCES

Agresti, A. (1996). *An Introduction to Categorical Data Analysis*. New York, NY: Wiley.

Berglund, P. (2011). "An Overview of Survival Analysis using Complex Sample Data," *Paper presented at the SAS Global Forum*. Las Vegas, NV, April 4 – 7. Available online at: http://support.sas.com/resources/papers/proceedings11/338-2011.pdf

Brown, L., Cai, T., and DasGupta, A. (2001). "Interval Estimation for a Binomial Proportion," *Statistical Scienc*e, **16**, pp. 101 – 133.

Clopper, C., and Pearson, E. (1934). "The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial," *Biometrika*, **26**, pp. 404 – 413.

Kish, L. (1965). *Survey Sampling*. New York, NY: Wiley.

Korn, E., and Graubard, B. (1998). "Confidence Intervals for Proportions with Small Expected Number of Positive Counts Estimated from Survey Data," *Survey Methodology*, **24**, pp. 193 – 201.

Lewis, T. (2012). "Modeling Complex Survey Data." *Paper presented at the Midwest SAS Users Group (MWSUG) Conference*. Minneapolis, MN, September 16 – 18. Available online at: http://www.mwsug.org/proceedings/2012/SA/MWSUG-2012-SA07.pdf

Lewis, T. (2013). "Analyzing Continuous Variables from Complex Survey Data Using PROC SURVEYMEANS," *Paper presented at the Midwest SAS Users Group (MWSUG) Conference*. Columbus, OH, September 22 – 24.

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

Rao, J.N.K., and Scott, A. (1981). "The Analysis of Categorical Data from Complex Sample Surveys: Chi-Squared Tests for Goodness of Fit and Independence in Two-Way Tables," *Journal of the American Statistical Association*, **76**, pp, 221 – 230.

Rao, J.N.K., and Scott, A. (1984). "On Chi-Squared Tests for Multiway Contingency Tables with Cell Proportions Estimated from Survey Data," *Annals of Statistics*, **12**, pp. 46 – 60.

Rust, K., and Hsu, V. (2007). "Confidence Intervals for Statistics for Categorical Variables from Complex Samples," *Proceedings of the 2007 Joint Statistical Meetings*, Salt Lake City, UT.

Thomas, D., and Rao, J.N.K. (1987). "Small-Sample Comparisons of Level and Power for Simple Goodness-of-Fit Statistics Under Cluster Sampling," *Journal of the American Statistical Association*, **82**, pp. 630 – 636.

Vollset, S. (1993). "Confidence Intervals for a Binomial Proportion," *Statistics in Medicine*, **12**, pp. 809 – 824.

Wilson, E. (1927). "Probable Inference, the Law of Succession, and Statistical Inference," *Journal of the American Statistical Association*, **22**, pp. 209 – 212.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Taylor Lewis
Joint Program in Survey Methodology (JPSM)
1218 LeFrak Hall
University of Maryland
College Park, MD 20742
Email: tlewis@survey.umd.edu