Analytics Recommendations for HINTS-FDA, Cycle 2

November, 2017

Table of Contents

Overview of HINTS

The Health Information National Trends Survey (HINTS) is a nationally-representative survey which has been administered every few years by the National Cancer Institute since 2003. The HINTS target population is all adults aged 18 or older in the civilian non-institutionalized population of the United States. The HINTS program collects data on the American public's need for, access to, and use of health-related information and health-related behaviors, perceptions and knowledge. (Hesse, et al., 2006; Nelson, et al., 2004). Previous iterations include HINTS 1 (2003), HINTS 2 (2005), HINTS 3 (2007/2008), HINTS 4, Cycle 1 (2011/2012), HINTS 4, Cycle 2 (2012/2013), HINTS 4, Cycle 3 (Late 2013), HINTS 4, Cycle 4 (2014), and HINTS-FDA, Cycle 1 (2015).

HINTS-FDA, Cycle 2

The first HINTS-FDA administration was conducted from May 2015 through September 2015, with a second administration conducted from January 2017 through May 2017. The second administration isthe focus of this report. HINTS-FDA, Cycle 2 was a special round of HINTS data collection conducted by the National Cancer Institute (NCI) in partnership with the Food and Drug Administration (FDA) tocombine the traditional HINTS topics of health communication, cancer knowledge, and cancer riskbehaviors with an assessment of the public's risk perceptions about new tobacco products, perceptionsof tobacco product harm, and tobacco product claims in addition to other FDA topics. HINTS-FDA, Cycle 2 was conducted by mail using a protocol similar to that used in HINTS-FDA, Cycle 1 with a goal ofobtaining 1,600 completed questionnaires. For more extensive background about the HINTSprogram and previous data collection efforts, see Finney Rutten et al. (2012).

Methodology

Data collection for HINTS-FDA, Cycle 2 was initiated in January 2017 and concluded in May 2017. HINTS-FDA, Cycle 2 was a self-administered mailed questionnaire. Because of the unique nature ofthe HINTS-FDA, Cycle 2 instrument and the specific goals of FDA, the regular sampling strategy of HINTS was altered in an effort to include more current and former smokers in the study. Additional adjustments were also made to the sampling strategy of HINTS-FDA, Cycle 1. Unlike HINTS-FDA, Cycle 1, which used adjusted county-level estimates of current smokers from the 2003 Behavioral Risk Factor Surveillance System (BRFSS), HINTS-FDA, Cycle 2 used county-level small area estimates of current smokers from the 2010-2011 Current Population Survey (TUS-CPS) to define the sampling strata. The high and the medium-high strata were then oversampled to increase the yield of current smokers. There was also a change in the sample alllocation across the sampling strata, as a change in the degree of differential sampling across the strata compared to the first FDA cycle was designed to yield a relatively larger number of expected smokers in terms of effective sample size (taking into account the design effect due to differential sampling).The sampling frame consisted of a database of addresses used by Marketing Systems Group (MSG) to provide random samples of addresses. All non-vacant residential addresses in the United States present on the MSG database, including post office (P.O.) boxes, throwbacks (i.e., street addresses for which mail is redirected by the United States Postal Service to a specified P.O. box), and seasonal addresses, were subject to sampling. A total of four mailings were sent out as part of HINTS-FDA, Cycle 2. The mailing protocol followed a modified Dillman approach (Dillman, et. al., 2009) with a total of four mailings: an initial mailing, a reminder postcard, and two follow-up mailings. All households in the sample received the first mailing and reminder postcard, while only non-responding households received the subsequent survey mailings. Most households received one survey per mailing (in English), while households that were flagged as potentially Spanish-speaking received two surveys per mailing (one English and one Spanish). The second-stage of sampling consisted of selecting one adult within each sampled household. In keeping with HINTS-FDA, Cycle 1, data collection for HINTS-FDA, Cycle 2 implemented the Next Birthday Method to select the one adult in the household. Questions were included on the survey instrument to assist the household in selecting the adult in the household having the next birthday. A $2 monetary incentive was included with the survey to encourage participation. Refer to the HINTS-FDA, Cycle 2 Methodology Report for more extensive information about the sampling procedures.

Sample Size and Response Rates

The final HINTS-FDA, Cycle 2 sample consists of 1,736 respondents. Note that 60 of these respondents were considered partial completers who did not answer the entire survey. A questionnaire was considered to be complete if at least 80% of Sections A and C were answered. A questionnaire was considered to be partially complete if 50% to 79% of the questions were answered in Sections A and C. Household response rates were calculated using the American Association for Public Opinion Research response rate 2 (RR2) formula. The overall household response rate using the Next Birthday method was 30.60%.

Analyzing HINTS Data

If you are solely interested in calculating point estimates (means, proportions etc.), either weighted or unweighted, you can use programs including SAS, SPSS, STATA and Systat. If you plan on doing inferential statistical testing using the data (i.e., anything that involves calculating a p value or confidence interval), it is important that you utilize a statistical program that can incorporate the replicate weights that are included in the HINTS database. The issue is that the standard errors in your analyses will most likely be underestimated if you don't incorporate the jackknife replicate weights; therefore, your p-values will be smaller than they "should" be, your tests will be more liberal, and you are more likely to make a type I error. Statistical programs like SUDAAN, STATA, SAS and Wesvar can incorporate the replicate weights found in the HINTS database. Currently, SPSS is not able to incorporate these replicate weights.

Note that analyses of HINTS variables that contain a large number of valid responses usually produce reliable estimates, but analyses of variables with a small number of valid responses may yield unreliable estimates, as indicated by their large variances. The analyst should pay particular attention to the standard error and coefficient of variation (relative standard error) for estimates of means, proportions, and totals, and the analyst should report these when writing up results. It is important that the analyst realizes that small sample sizes for particular analyses will tend to result in unstable estimates.

Important Analytic Variables in the Database

Note: Refer to the HINTS-FDA, Cycle 2 Methodology Report for more information regarding the weighting and stratification variables listed below.

PERSON_FINWT0: Final sample weight used to calculate population estimates. Note that estimates from the 2014 American Community Survey (ACS) of the US Census Bureau were used to calibrate the HINTS-FDA control totals with the following variables: Age, gender, education, marital status, race, ethnicity, and census region. In addition, variables from the 2015 National Health Interview Survey (NHIS) were used to calibrate HINTS-FDA, Cycle 2 data control totals regarding: Percent with health insurance and percent ever had cancer.

PERSON_FINWT1 THROUGH PERSON_FINWT50: Fifty replicate weights that can be used to calculate accurate standard error of estimates using the jackknife replication method. More information about how these weights were created can be found in the “HINTS-FDA, Cycle 2 Methodology Report” included in the data download, or see Korn and Graubard (1999).

STRATUM: This variable codes for whether the respondent was in the Low, Medium-Low, Medium-High or High smoking rate sampling stratum.

HIGHSPANLI: This variable codes for whether the respondent was in the High Spanish Linguistically Isolated stratum (Yes or No).

HISPSURNAME: This variable codes for whether there was a Hispanic surname match for this respondent (Yes or No).

HISP_HH: This variable codes for households identified as Hispanic by either being in a high linguistically isolated strata, or having a Hispanic surname match, or both.

APP_REGION: This variable codes for Appalachia subregion.

LANGUAGE_FLAG: This variable codes for language the survey was completed in (English or Spanish).

QDISP: This variable codes for whether the survey returned by the respondent was considered Complete or Partial Complete. A complete questionnaire was defined as any questionnaire with at least 80% of the required questions answered in Sections A and C. A partial complete was defined as when between 50% and 79% of the questions were answered in Sections A and C. There were 143 partially complete questionnaires. Sixty-three questionnaires with fewer than 50% of the required questions answered in Sections A and C were coded as incompletely-filled out and discarded.

INCOMERANGES_IMP: This is the income variable (INCOMERANGES) imputed for missing data. To impute for missing items, PROC HOTDECK from the SUDAAN statistical software was used. PROC HOTDECK uses the Cox-Iannacchione Weighted Sequential Hot Deck imputation method as described by Cox (1980). The following variables were used as imputation classes given their strong association with the income variable: Education (O6), Race/Ethnicity (RaceEthn), Do you currently rent or own your house? (O15), How well do you speak English? (O9), and Were you born in the United States? (O7).

Denominator Degrees of Freedom (DDF)

The HINTS-FDA, Cycle 2 database contains a set of 50 replicate weights to compute accurate standard errors for statistical testing procedures. These replicate weights were created using a jackknife minus one replication method; when analyzing one iteration of HINTS data, the proper denominator degrees of freedom (ddf) is 49. Thus, analysts who are only using the HINTS-FDA, Cycle 2 data should use 49 ddf in their statistical models. HINTS statistical analyses that involve more than one iteration of data will typically utilize a set of 50*k replicate weights, where they can be viewed as being created using a stratified jackknife method with k as the number of strata, and 49*k as the appropriate ddf. Analysts who were merging two iterations of data and making comparisons should adjust the ddf to be 98 (49*2) etc.

References

Cox, B. G. (1980). “The Weighted Sequential Hot Deck Imputation Procedure”. Proceedings of the American Statistical Association, Section on Survey Research Methods.

Dillman, D.A., Smyth, J.D., and Christian, L.M. (2009). *Internet, mail, and mixed-mode surveys: The tailored design method.* Hoboken, NJ: John Wiley & Sons.

Finney Rutten, L. J., Davis, T., Beckjord, E. B., Blake, K., Moser, R. P., & Moser, R. P. (2012) Picking Up the Pace: Changes in Method and Frame for the Health Information National Trends Survey (2011 – 2014). Journal of Health Communication, 17 (8), 979-989.

Hesse, B. W., Moser, R. P., Rutten, L. J., & Kreps, G. L. (2006). The health information national trends survey: research from the baseline. *J Health Commun, 11 Suppl 1*, vii-xvi.

Korn, E. L., & Graubard, B. I. (1999). Analysis of health surveys. New York: John Wiley & Sons.

Nelson, D. E., Kreps, G. L., Hesse, B. W., Croyle, R. T., Willis, G., Arora, N. K., et al. (2004). The Health Information National Trends Survey (HINTS): development, design, and dissemination. *J Health Commun, 9*(5), 443-460; discussion 481-444.

Appendix

The following appendices A through C provide some coding examples using SAS, SUDAAN, and STATA for common types of statistical analyses using HINTS-FDA, Cycle 2 data. These examples will incorporate both the final sample weight (to get population estimates) and the set of 50 jackknife replicate weights to get the proper standard error. Although these examples specifically use HINTS-FDA, Cycle 2 data, the concepts used here are generally applicable to other types of analyses. We will consider an analysis that includes gender, education level (edu) and two questions that are specific to the HINTS-FDA, Cycle 2 data: seekhealthinfo & friendsusetobacco.

Appendices D and E provide SAS code to combine HINTS FDA, Cycle 1 and HINTS FDA, Cycle 2 survey iterations and HINTS FDA, Cycle 2 and HINTS 5 Cycle 1 survey iterations. The provided code will generate one final sample weight for population point estimates and 100 replicate weights to compute standard errors.

- **Appendix A:** Analyzing data using SAS
- **Appendix B:** Analyzing data using SUDAAN
- **Appendix C:** Analyzing data using STATA
- **Appendix D:** Merging HINTS FDA, Cycle 1 and HINTS FDA, Cycle 2 using SAS
- **Appendix E:** Merging HINTS 5, Cycle 1 and HINTS FDA, Cycle 2 using SAS

Appendix A: Analyzing data using SAS

This section gives some SAS (Version 9.3 and higher) coding examples for common types of statistical analyses using HINTS-FDA, Cycle 2 data. We begin by doing data management of the HINTS-FDA, Cycle 2 data in a SAS DATA step. We first decided to exclude all "Missing data (Not Ascertained)" and "Multiple responses selected in error" responses from the analyses. By setting these values to missing (.), SAS will exclude these responses from procedures where these variables are specifically accessed. For logistic regression modeling within the PROC SURVEYLOGISTIC procedure, SAS expects the response variable to be dichotomous with values (0, 1), so this variable will also be recoded at this point. It is better to use dummy variables instead of categorical variables in SAS survey procedures, such as PROC SURVEYREG. We use dummy variables for gender and education level in both PROC SURVEYLOGISTIC and PROC SURVEYREG procedures. When recoding existing variables, it is generally recommended to create new variables, rather than over-writing the existing variables. Note: New variables should always be compared to original source variables in a SAS PROC FREQ procedure to verify proper coding.

```
options fmtsearch=(HFDA2);  *This is used to call up the formats;
*substitute your library name in the parentheses;

proc format;  *First create some temporary formats;

Value Genderf
1 = "Male"
2 = "Female";

Value Educationf
1 = "Less than high school"
2 = "12 years or completed high school"
3 = "Some college"
4 = "College graduate or higher";

value seekhealthinfof
1 = "Yes"
0 = "No";
run;
```

```
data hints_fda2;
set HFDA2.hints_fda2_public;

/*Recode negative values to missing*/
if Selfgender = 1 then gender =  1;
if Selfgender = 2 then gender =  2;
if selfgender<0 then gender = .;

/*Recode education into four levels, and negative values to missing*/
if education in (1, 2) then edu = 1;
if education = 3 then edu = 2;
if education in (4, 5) then edu = 3;
if education in (6, 7) then edu = 4;
if education <0 then edu = .;
```

```
/*Recode seekhealthinfo to 0-1 format for proc rlogist procedure,
and negative values to missing */
if seekhealthinfo = 2 then seekhealthinfo = 0;
if seekhealthinfo<0 then seekhealthinfo = .;

/*Recode negative values to missing for proc regress procedure*/
if FriendsUseTobacco<0 then FriendsUseTobacco=.;

/*Create dummy variables for proc surveylogistic and proc surveyreg
procedures*/
if gender = 1 then
Female = 0;
else if gender = 2 then
Female = 1;

if edu = 1 then do;
HighSchool = 0;
SomeCollege = 0;
CollegeorMore = 0;
end;

else if edu = 2 then do;
HighSchool = 1;
SomeCollege = 0;
CollegeorMore = 0;
end;

else if edu = 3 then do;
HighSchool = 0;
SomeCollege = 1;
CollegeorMore = 0;
end;

else if edu = 4 then do;
HighSchool = 0;
SomeCollege = 0;
CollegeorMore = 1;
end;

/*Apply formats to recoded variables */
format gender genderf. edu educationf. seekhealthinfo
seekhealthinfof.; run;
```

Proc Surveyfreq procedure

We are now ready to begin using SAS 9.3 to examine the relationships among these variables. Using **PROC SURVEYFREQ**, we will first generate a cross-frequency table of education by gender, along with a (Wald) Chi-squared test of independence. Note the syntax of the overall sample weight, PERSON_FINWT0, and those of the jackknife replicate weights, PERSON_FINWT1—PERSON_FINWT50. The jackknife adjustment factor for each replicate weight is 0.98. This syntax is consistent for all procedures. Other data sets that incorporate replicate weight jackknife designs will follow a similar syntax.

```
proc surveyfreq data = hints_fda2 varmethod =jackknife;
      weight person_finwt0;
      repweights person_finwt1-person_finwt50 / df = 49 jkcoefs = 0.98;
      tables edu*gender / row col wchisq;
run;
```

The *tables* statement defines the frequencies that should be generated. Stand-alone variables listed here result in one-way frequencies, while a “*” between variables will define cross-frequencies. The *row* option produces row percentages and standard errors, allowing us to view stratified percentages. Similarly, the *col* option produces column percentages and standard errors, allowing us to view stratified percentages. The option *wchisq* requests Wald chi-square test for independence. Other tests and statistics are also available; see the SAS 9.3 Product Documentation Site for more information.

For the purposes of computing appropriate degrees of freedom for the estimator of the HINTS-FDA, Cycle 2 differences, we can assume, as an approximation, that the sample is a simple random sample of size 50 (corresponding to the 50 replicates: each replicate provides a ‘pseudo sample unit’) from a normal distribution. The denominator degrees of freedom (df) is equal to 49*k, where k is the number of iterations of data used in this analysis.

Variance Estimation	
Method	Jackknife
Replicate Weights	HINTS_FDA2
Number of Replicates	50

Table Education by Gender

edu	**gender**	**Frequency**	**Percent**	**Std Err of Percent**	**Row Percent**	**Std Err of Row Percent**	**Column Percent**	**Std Err of Col Percent**
Less than high school	Male	42	3.4179	0.6938	59.2541	7.660	7.0468	1.4304
	Female	34	2.3503	0.5454	40.7459	7.660	4.5640	1.0644
	Total	76	5.7682	0.8748	100			
12 years or completed high school	Male	119	11.8809	1.4129	47.9615	3.821	24.4951	2.9089
	Female	184	12.8908	1.3550	52.0385	3.821	25.0321	2.6734
	Total	303	24.7716	2.0256	100			
Some college	Male	176	16.1196	1.0768	46.8317		33.2342	2.1726
	Female	253	18.3009	2.2011	53.1683	3.902	35.5373	4.1380
	Total	429	34.4202	2.2075	100			

edu	gender	Frequency	Percent	Std Err of Percent	Row Percent	Std Err of Row Percent	Column Percent	Std Err of Col Percent
College graduate or higher	Male	334	17.0846	1.0923	48.7576	2.810	35.2239	2.1056
	Female	441	17.9553	0.9824	51.2424	2.810	34.8667	1.9259
	Total	775	35.0399	0.6631	100			
Total	Male	671	48.5029	0.7749			10	
	Female	912	51.4971	0.7749			10	
	Total	1583	10					

Frequency Missing = 153

Wald Chi-Square Test	
Chi-Square	1.9255
F Value	0.6418
Num DF	3
Den DF	49
Pr > F	0.5918
Adj F Value	0.6156
Num DF	3
Den DF	47
Pr > Adj F	0.6083

Sample Size = 1,583

The weighted percentages above show that a similar proportion of women have at least a college degree compared to men, 17.96% vs. 17.08%. The Chi-squared test of independence indicates that there is not a significant difference between these the educational distribution in these two groups (p-value > 0.05).

Logistic Regression

This example demonstrates a multivariable logistic regression model using **PROC SURVEYLOGISTIC**; recall that the response should be a dichotomous 0-1variable.

```
/*Multivariable logistic regression of gender and education on
SeekHealthInfo*/
proc surveylogistic data= hints_fda2 varmethod=jackknife;
 weight person_finwt0;
 repweights person_finwt1-person_finwt50 / df=49 jkcoefs=0.98;
 model seekhealthinfo (descending) = Female HighSchool SomeCollege
 CollegeorMore / tech=newton xconv=1e-8;
 contrast 'Overall model' intercept 1, Female 1,
 HighSchool 1,
 SomeCollege 1, CollegeorMore 1;
 contrast 'Overall model minus intercept' Female 1, HighSchool 1,
 SomeCollege 1,
 CollegeorMore 1;
 contrast 'Gender' Female 1;
 contrast 'Education overall' HighSchool 1, SomeCollege 1, CollegeorMore 1;
run;
```

The response variable should be on the left hand side (LHS) of the equal sign in the model statement, while all covariates should be listed on the right hand side (RHS). The *descending* option requests the probability of seekhealthinfo="Yes" to be modeled. The "Male" is the reference group for gender effect while "Less than high school" is the reference group for education level effect. The option *tech=newton* requests the Newton-Raphson algorithm. The option xconv=1e-8 helps to avoid early termination of the iteration.

Variance Estimation	
Method	Jackknife
Replicate Weights	HINTS_FDA2
Number of Replicates	50

Analysis of Maximum Likelihood Estimates

Parameter	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept	1	0.4549	0.4098	1.2324	0.2670
Female	1	0.0804	0.2300	0.1222	0.7267
HighSchool	1	0.4603	0.4466	1.0621	0.3027
SomeCollege	1	1.3166	0.4604	8.1759	0.0042
CollegeorMore	1	1.5294	0.4330	12.4747	0.0004

Odds Ratio Estimates

Effect	**Point Estimate**	**95% Wald Confidence Limits**	
Female	1.084	0.690	1.701
HighSchool	1.585	0.660	3.803
SomeCollege	3.731	1.513	9.198
CollegeorMore	4.615	1.975	10.784

Contrast Test Results

Contrast	**DF**	**Wald Chi-Square**	**Pr > ChiSq**
Overall model	5	287.6170	<.000
Overall model minus intercept	4	29.9713	<.000
Gender	1	0.1222	0.7267
Education overall	3	28.4504	<.000

To identify levels/variables that display a significant difference in response, the rule of thumb is to examine odds ratios where the confidence interval does not contain 1 (by default, SAS will use alpha=.05 to determine statistical significance; this value can be changed by the user using code). However, significance may also be garnered from the test of whether the associated beta parameter is equal to 0 (see first regression table above). According to this model, those with some college or a college degree or more appear to be statistically more inclined to search for health information. Gender is not a significant variable.

Linear Regression

This example demonstrates a multivariable linear regression model using **PROC SURVEYREG**; recall that the response should be a continuous variable. For the purposes of this example, we decided to use an outcome with six levels as a continuous variable (FriendsUseTobacco). Note that higher values on FriendsUseTobacco indicate the more friends of the participant using tobacco.

```
 /*Multivariable linear regression of gender and education on
FriendsUseTobacco*/
proc surveyreg data= hints_fda2 varmethod=jackknife;
weight person_finwt0;
repweights person_finwt1-person_finwt50 / df=49 jkcoefs=0.98;
model FriendsUseTobacco = Female HighSchool SomeCollege CollegeorMore;
contrast 'Overall model' intercept 1,
Female 1,
HighSchool 1, SomeCollege 1, CollegeorMore 1;
contrast 'Overall model minus intercept' Female 1,
HighSchool 1,
SomeCollege 1, CollegeorMore 1;
contrast 'Gender' Female 1;
contrast 'Education overall' HighSchool 1,
SomeCollege 1,
CollegeorMore 1;
run;
```

Variance Estimation	
Method	Jackknife
Replicate Weights	HINTS_FDA2
Number of Replicates	50

Analysis of Contrasts

Contrast	**Num DF**	**F Value**	**Pr > F**
Overall model	5	75.43	<.000
Overall model minus intercept	4	8.79	<.000
Gender	1	0.00	0.9619
Education overall	3	11.27	<.000

NOTE: The denominator degrees of freedom for the F tests is 49.

From the above table, we can see that only education is associated with the number of friends who use tobacco, adjusting for all variables in the model.

Estimated Regression of Coefficients

Parameter	**Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	2.1961185	0.30615510	7.17	<.0001
Female	0.0087562	0.18228031	0.05	0.9619
HighSchool	-0.8086069	0.35245719	-2.29	0.0261
SomeCollege	-0.8144547	0.39564762	-2.06	0.0449
CollegeorMore	-1.4470505	0.30469513	-4.75	<.0001

NOTE: The denominator degrees of freedom for the t-tests is 49.

From the above table, it can be seen that, those with at least a high school education have a significantly inverse linear association with number of friends who use tobacco (i.e., less friends using tobacco), controlling for all variables in the model. We don’t interpret gender because it is non-significant.

Appendix B: Analyzing data using SUDAAN

This section gives some SUDAAN (Version 11.0 and higher) coding examples for common types of statistical analyses using HINTS-FDA, Cycle 2 data. We begin by doing data management of the HINTS-FDA, Cycle 2 data in a SAS DATA step. We first decided to exclude all “Missing data (Not Ascertained)” and “Multiple responses selected in error” responses from the analyses. By setting these values to missing (.), SAS will exclude these responses from procedures where these variables are specifically accessed. For logistic regression modeling within the PROC RLOGIST procedure, SUDAAN expects the response variable to be dichotomous with values (0, 1), so this variable will also be recoded at this point. When recoding existing variables, it is generally recommended to create new variables of rather than over-writing the existing variables. Note: New variables should always be compared to original source variables in a SAS PROC FREQ procedure to verify proper coding.

```
 proc format;  *First create some temporary formats;

 Value Genderf
 1 = "Male"
 2 = "Female";

 Value Educationf
 1 = "Less than high school"
 2 = "12 years or completed high school"
 3 = "Some college"
 4 = "College graduate or higher";

 value seekhealthinfof
 1 = "Yes"
 0 = "No";

 run;


 data hints_fda2; /*CREATE A TEMPORARY DATA FILE FOR
 ANALYSIS*/ set hints.hints_fda2_public;

/*Recode negative values to missing and create new gender variable*/
if selfgender = 1 then gender = 1;
if selfgender = 2 then gender = 2;
if selfgender in (-9, -6) then gender = .;

/*Recode education into four levels, and negative values to missing*/
if education in (1, 2) then edu = 1;
if education = 3 then edu = 2;
if education in (4, 5) then edu = 3; if education in (6, 7) then edu = 4;
if education < 0 then edu = .;

/*Recode seekhealthinfo to 0-1 format for proc rlogist procedure, and negative
values to missing */
if seekhealthinfo = 2 then seekhealthinfo = 0;
if seekhealthinfo<0 then seekhealthinfo = .;

/*Recode negative values to missing for proc regress procedure*/
if FriendsUseTobacco<0 then FriendsUseTobacco=.;
```

```
/*Apply formats to recoded variables */
format gender genderf. edu educationf. seekhealthinfo seekhealthinfof.;
run;
```

We are now ready to begin using SUDAAN to examine the relationships among these variables. Using **proc crosstab**, we will first generate a cross-frequency table of education and gender, along with a (Wald) Chi-squared test of independence. Note the syntax of the overall sample weight, PERSON_FINWT0, and those of the jackknife replicate weights, PERSON_FINWT1—PERSONFINWT50. The jackknife adjustment factor for each replicate weight is 0.98. This syntax is consistent for all procedures. Other data sets that incorporate replicate weight jackknife designs will follow a similar syntax.

```
proc crosstab data= hints_fda2 design=jackknife ddf = 49;
weight person_finwt0;
jackwgts person_finwt1-person_finwt50 / adjjack=.98;
class gender edu;
tables edu*gender;
test chisq;
run;
```

Since this procedure is mainly for categorical variables, each variable should be specified as such by inclusion in the class statement (which is ubiquitous in all SUDAAN procedures). The *tables* statement defines the frequencies that should be generated. Stand-alone variables listed here result in one-way frequencies, while a "*" between variables will define cross-frequencies. In general, the PROC CROSSTAB procedure may be used to investigate n-way variable frequencies, along with their relationships. This is accomplished by the *test* statement, which defines various types of independence tests: here a Chi-Squared test is implemented. Other tests and statistics are also available; see the SUDAAN site link for more information.

The HINTS-FDA, Cycle 2 database for a single iteration contains a set of 50 replicate weights to compute accurate standard errors for statistical testing procedures. These replicate weights were created using a jackknife minus one replication method. Thus, the proper denominator degrees of freedom (ddf) should be 49 when one iteration of HINTS data is being analyzed. Thus, analysts who are only using the HINTS-FDA, Cycle 2 data should use 49 ddf in their statistical models.

HINTS-FDA, Cycle 2 databases with more than one iteration of data will contain a set of 50*k replicate weights, where they can be viewed as being created using a stratified jackknife method with k as the number of strata and 49*k as the appropriate ddf. Analysts who were merging two iterations of data and making comparisons these should adjust the ddf to be 98 (49*2) etc.

Variance Estimation Method: Replicate Weight Jackknife
By: EDU, GENDER

What is the highest grade or level of schooling you completed?		Are you male or female? Total	Male	Female
Total	**Sample Size**	1583	671	912
	Col Percent	100	100	100
	Row Percent	100	48.50	51.50
Less than HS	**Sample Size**	76	42	34
	Col Percent	5.77	7.05	4.56
	Row Percent	100	59.25	40.75
12 years or completed HS	**Sample Size**	303	119	184
	Col Percent	24.77	24.50	25.03
	Row Percent	100	47.96	52.04
Some college	**Sample Size**	429	176	253
	Col Percent	34.42	33.23	35.54
	Row Percent	100	46.83	53.17
College graduate or higher	**Sample Size**	775	334	441
	Col Percent	35.04	35.22	34.87
	Row Percent	100	48.76	51.24

Variance Estimation Method: Replicate Weight Jackknife
Chi Square Test of Independence for EDU and GENDER

ChiSq	0.6418
P-value for ChiSq	0.5918
Degress of Freedom ChiSq	3

Logistic Regression

This example demonstrates a multivariable logistic regression model using **PROC RLOGIST** (*RLOGIST* is used to differentiate it from the SAS procedure, PROC LOGISTIC, and is used with SAS-callable SUDAAN); recall that the response should be a dichotomous 0-1 variable.

```
/*Multivariable logistic regression of gender and education on
Seekhealthinfo*/
proc rlogist data = hints_fda2 design = jackknife ddf = 49;
weight person_finwt0;
jackwgts person_finwt1-person_finwt50 / adjjack =0.98;
class gender edu;
model seekhealthinfo = gender edu;
reflev gender=1 edu=1;
run;
```

The response variable should be on the left hand side (LHS) of the equal sign in the model statement, while all covariates should be listed on the right hand side (RHS). Categorical variables should also be included in the class statement. By default, the reference level of each categorical variable is that of the highest numeric level. In this example, males and those with less than a high school education were changed to be the reference values for gender and education, respectively, by using the reflev statement to explicitly define another reference level.

Variance Estimation Method: Replicate Weight Jackknife
Working Correlations: Independent
Link Function: Logit
Response variable SEEKHEALTHINFO: A1. Have you ever looked for information about cancer from any source?
by: Independent Variables and Effects.

Independent variables and effects	Beta Coeff.	SE Beta	T-test B=0	P-value T-Test B=0
Intercept	0.45	0.41	1.11	0.2724
Gender				
Male	0.00	0.00	-	-
Female	0.08	0.23	0.35	0.7282
Education Level				
Less than HS	0.00	0.00	-	-
12 years or HS completed	0.46	0.45	1.03	0.3078
Some College	1.32	0.46	2.86	0.0062
College graduate or higher	1.53	0.43	3.53	0.0009

Contrast Test Results

Contrast	Degrees of Freedom	Wald F	P-value Wald Chi-Sq
Overall Model	5	57.52	<.0001
Model minus intercept	4	7.49	0.0001
Intercept	--	--	--
Gender	1	0.12	0.7282
Edu	3	9.48	<.0001

Odds Ratio Estimates

Independent variables and effects	Odds Ratio	Lower 95% Limit OR	Upper 95% Limit OR
Intercept	1.58	0.69	3.59
Gender			
Male	1.00	1.00	1.00
Female	1.08	0.68	1.72
Education Level			
Less than HS	1.00	1.00	1.00
12 years or HS completed	1.58	0.65	3.89
Some College	3.73	1.48	9.41
College graduate or higher	4.62	1.93	11.02

To identify levels/variables that display a significant difference in response, the rule of thumb is to examine odds ratios where the confidence interval does not contain 1 (by default, SUDAAN will use alpha=.05 to determine statistical significance; this value can be changed by the user using code). However, significance may also be garnered from the test of whether the associated beta parameter is equal to 0 (see first regression table above). According to this model, those with some college or a college degree or more appear to be statistically more inclined to search for health information. Gender is not a significant variable (p-value > 0.05).

Linear Regression

This example demonstrates a multivariable linear regression model using **PROC REGRESS** (REGRESS is used to differentiate it from the SAS procedure, PROC REG, and is used with SAS-callable SUDAAN); recall that the response should be a continuous variable. For the purposes of this example, we decided to use an outcome with six levels as a continuous variable (FRIENDSUSETOBACCO). Note that higher values on FRIENDSUSETOBACCO indicate the more friends of the participant using tobacco.

The response variable should be on the left hand side (LHS) of the equal sign in the model statement, while all covariates should be listed on the right hand side (RHS). Categorical variables should also be included in the class statement. By default, the reference level of each categorical variable is that of the highest numeric level. In this example, males and those with less than a high school education were changed to be the reference values for gender and education, respectively, by using the reflev statement to explicitly define another reference level.

```
/*Multivariable linear regression of gender and education on GeneralHealth*/
proc regress data = hints_fda2 design = jackknife ddf = 49;
weight person_finwt0;
jackwgts person_finwt1-person_finwt50 / adjjack = 0.98;
class gender edu;
model friendsusetobacco = gender edu ;
reflev gender=1 edu=1;
run;
```

Variance Estimation Method: Replicate Weight Jackknife
Working Correlations: Independent
Link Function: Identity
Response variable FRIENDSUSETOBACCO: C19. Of the five closest friends or acquaintances that you

spend time with, how many of them use tobacco?
by: Contrast.

Contrast	Degrees of Freedom	Wald F	P-value Wald F
Overall Model	5	75.43	0.0000
Model minus intercept	4	8.97	0.0000
Intercept	--	--	--
Gender	1	0.00	0.9619
Edu	3	11.27	0.0000

From the above table, we can see that Gender is not associated with the outcome, but Edu is associated, adjusting for all variables in the model.

Variance Estimation Method: Replicate Weight Jackknife
Working Correlations: Independent
Link Function: Identity
Response variable FRIENDSUSETOBACCO: C19. Of the five closest friends or acquaintances that you spend time with, how many of them use tobacco?

Independent variables and effects	Beta Coeff.	SE Beta	T-test B=0	P-value T-Test B=0
Intercept	2.20	0.31	7.17	0.0000
Gender				
Male	0.00	0.00	-	-
Female	0.01	0.18	0.05	0.9619
Education Level				
Less than HS	0.00	0.00	-	-
12 years or HS completed	-0.81	0.35	-2.29	0.0261
Some College	-0.81	0.40	-2.06	0.0449
College graduate or higher	-1.45	0.30	-4.75	0.0000

From the above table, it can be seen that, compared to those respondents with Less than a High School education, those with a 12 years or HS completed have a significantly negative linear association with number of friends who use tobacco (i.e., fewer friends using tobacco), controlling for all variables in the model. This association also applies to those with Some College and College Degree or Higher. We don’t interpret the Gender variable because it is non-significant.

Appendix C: Analyzing data using STATA

This section gives some Stata (Version 10.0 and higher) coding examples for common types of statistical analyses using HINTS-FDA, Cycle 2 data. We begin by doing data management of the HINTS-FDA, Cycle 2 data. We first decided to exclude all “Missing data (Not Ascertained)”, “Multiple responses selected in error”, “Question answered in error (Commission Error)” and “Inapplicable, coded 2 in SeekHealthInfo” responses from the analyses. By setting these values to missing (.), Stata will exclude these responses from analysis commands where these variables are specifically accessed. For logistic regression modeling within the **svy: logit** command, Stata expects the response variable to be dichotomous with values (0, 1), so this variable will also be recoded at this point. When recoding existing variables, it is generally recommended to create new variables of rather than over-writing the existing variables. Note: New variables should always be compared to original source variables in a Stata **tabulate** command to verify proper coding.

```
use “file path\hints_fda2_public.dta”
* Recode negative values to missing

recode selfgender (1=1 "Male") (2=2 "Female") (nonmissing=.), generate(gender)

label variable gender "Gender"

* Recode education into four levels, and negative values to missing

recode education (1/2=1 "Less than high school") (3=2 "12 years or completed high
school") (4/5=3 "Some college") (6/7=4 "College graduate or higher")
(nonmissing=.), generate(edu)

label variable edu "Education"



* Recode seekhealthinfo to 0-1 format, and negative values to missing for

svy: logit

replace seekhealthinfo = 0 if seekhealthinfo == 2

replace seekhealthinfo = . if seekhealthinfo == -9

label define seekhealthinfo2 0 "No" 1 "Yes"
label values seekhealthinfo seekhealthinfo2


* Recode negative values to missing for svy: regress

replace friendsusetobacco = . if friendsusetobacco == -5 | friendsusetobacco ==
-9
```

Declare survey design

Stata requires declaring the survey design for the data set globally before any analysis. The declared survey design will be applied to all future survey commands unless another survey design is declared. Other data sets that incorporate the final sample weight and the 50 jackknife replicate weights will utilize the same code.

```
* Declare survey design for the data set

svyset [pw=person_finwt0], jkrw(person_finwt1-person_finwt50,
multiplier(0.98)) vce(jack) mse
```

Cross-tabulation

```
* cross-tabulation

svy: tabulate edu gender, column row format(%8.5f) percent wald noadjust
```

The **svy: tabulate** command defines the frequencies that should be generated. Single variables listed in **svy: tabulate** results in one-way frequencies, while two variables will define cross-frequencies. The options **column** and **row** request column and row frequencies, respectively. The option **percent** requests the frequencies are displayed in percentage. The options **wald** and **noadjust** together request unadjusted Wald test for independence. Stata recommends default pearson test for independence. Other tests and statistics are also available; see the Stata website for more information: http://www.stata.com/

For the purposes of computing appropriate degrees of freedom for the estimator of the HINTS-FDA, Cycle 2 cycle differences, we can assume as an approximation that the sample is a simple random sample ofsize 50 (corresponding to the 50 replicates: each replicate provides a ‘pseudo sample unit’) from a normal distribution. The denominator degrees of freedom (df) is equal to 49*k, where k is the number of iterations of data used in this analysis. Stata uses the number of replicates minus one as the denominator degrees of freedom and does not provide the option for user to specify the denominator degrees of freedom.

```
Jknife *: for cell counts

Number of strata   =       1              Number of obs     =       1,583
                                          Population size   =  227,440,668
                                          Replications      =          50
                                          Design df         =          49
```

Education	Gender		Total
	Male	Female	
Less than HS	59.25413	40.74587	1.00E+02
	7.04682	4.56398	5.76823
12 years or HS completed	47.96153	52.03847	1.00E+02
	24.49511	25.03207	24.77163
Some college	46.83167	53.16833	1.00E+02
	33.23419	35.53727	34.42021
College graduate or higher	48.75757	51.24243	1.00E+02
	35.22387	34.86668	35.03993
Total	48.50294	51.49706	1.00E+02
	1.00E+02	1.00E+02	1.00E+02

Key: row percentage
column percentage

```
Wald (Pearson):
  Unadjusted   chi2(3)      =   1.9255
  Unadjusted   F(3, 49)     =   0.6418    P = 0.5918
  Adjusted     F(3, 47)     =   0.6156    P = 0.6083
```

Logistic Regression

This example demonstrates a multivariable logistic regression model using **svy: logit** (to get parameters) and **svy, or: logit** (to get odds ratios); recall that the response should be a dichotomous 0-1 variable.

```
* Define reference group for categorical variables for both svy: logit and
svy: regress

char gender [omit] 1

char edu [omit] 1


* Multivariable logistic regression of gender and education on seekhealthinfo

xi: svy: logit seekhealthinfo i.gender i.edu

test _Igender_2 _Iedu_2 _Iedu_3 _Iedu_4 _cons, nosvyadjust

test _Igender_2 _Iedu_2 _Iedu_3 _Iedu_4, nosvyadjust
test _Igender_2, nosvyadjust

test _Iedu_2 _Iedu_3 _Iedu_4, nosvyadjust

xi: svy, or: logit seekhealthinfo i.gender i.edu
```

The **char** command defines categorical variable with reference group. The “Male” is the reference group for gender effect while the “Less than high school” is the reference group for education level effect. These definitions will be applied to future commands until another **char** command re-defines the reference group. The xi command will create proper dummy variables for i.gender and i.edu variables in the analysis commands. The response variable should be the first variable in **svy: logit** command and be followed by all covariates. The **test** command tests the hypotheses about estimated parameters.

```
i.gender          _Igender_1-2        (naturally coded; _Igender_1 omitted)

i.edu             _Iedu_1-4            (naturally coded; _Iedu_1 omitted)

(running logit on estimation sample)


Jackknife replications (50)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+---  5
..................................................    50

Survey: Logistic regression

Number of strata   =        1              Number of obs     =          1,562
                                           Population size   =  224,841,160
                                           Replications      =             50
                                           Design df         =             49
                                           F(   4,     46)   =           7.03
                                           Prob > F          =         0.0002
```

seekcancer~o	Coef.	Jknife * Std. Err.	t	P>|t|	[95% Conf. Interval]	
_Igender_2	0.080404	0.230029	0.35	0.728	-0.38186	0.542665
_Iedu_2	0.460303	0.446638	1.03	0.308	-0.43725	1.357855
_Iedu_3	1.31657	0.460444	2.86	0.006	0.391274	2.241866
_Iedu_4	1.529369	0.433009	3.53	0.001	0.659204	2.399533
_cons	0.454874	0.409754	1.11	0.272	-0.36856	1.278306

```
  Unadjusted Wald test

 (1) [seekhealthinfo]_Igender_2 = 0
 (2)  [seekhealthinfo]_Iedu_2 = 0
 (3)  [seekhealthinfo]_Iedu_3 = 0
 (4)  [seekhealthinfo]_Iedu_4 = 0

 (5) [seekhealthinfo]_cons = 0

    F(  5,    49) =  57.52
         Prob > F =    0.0000

  Unadjusted Wald test

 (1) [seekhealthinfo]_Igender_2 = 0
 (2)  [seekhealthinfo]_Iedu_2 = 0
 (3)  [seekhealthinfo]_Iedu_3 = 0

 (4) [seekhealthinfo]_Iedu_4 = 0

    F(  4,    49) =  7.49
         Prob > F =    0.0001

  Unadjusted Wald test

( 1)  [seekhealthinfo]_Igender_2 = 0

    F(  1,    49) =   0.12
         Prob > F =    0.7282

  Unadjusted Wald test

 (1) [seekhealthinfo]_Iedu_2 = 0
 (2)  [seekhealthinfo]_Iedu_3 = 0

 (3) [seekhealthinfo]_Iedu_4 = 0

    F(  3,    49) =  9.58
         Prob > F =    0.0000

  i.gender          _Igender_1-2        (naturally coded; _Igender_1 omitted)

  i.edu              _Iedu_1-4            (naturally coded; _Iedu_1 omitted)

  (running logit on estimation sample)


  Jackknife replications (50)

  ----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5

  ..................................................    50


  Survey: Logistic regression
```

```
Number of strata   =        1              Number of obs     =       1,562
                                          Population size   =  224,841,160
                                           Replications      =          50
                                           Design df         =          49
                                           F(   4,     46)   =        7.03
                                           Prob > F          =      0.0002
```

seekhealthinfo	Odds Ratio	Jknife Std. Err.	t	P>|t|	[95% Conf. Interval]	
_Igender_2	1.083725	0.249289	0.35	0.728	0.682593	1.720587
_Iedu_2	1.584555	0.707722	1.03	0.308	0.645811	3.887845
_Iedu_3	3.730604	1.717733	2.86	0.006	1.478863	9.410879
_Iedu_4	4.615263	1.998451	3.53	0.001	1.933254	11.01803

To identify levels/variables that display a significant difference in response, the rule of thumb is to examine odds ratios where the confidence interval does not contain 1 (by default, Stata will use alpha=.05 to determine statistical significance; this value can be changed by the user using code). However, significance may also be garnered from the test of whether the associated beta parameter is equal to 0 (see first regression table above). According to this model, those with some college or a college degree or higher appear to be statistically more inclined to search for health information compared with those who did not graduate from high school, controlling for all other variables. Gender is not a significant variable.

Linear Regression

This example demonstrates a multivariable linear regression model using **svy: regress**; recall that the response should be a continuous variable. For the purposes of this example, we decided to use an outcome with six levels as a continuous variable (friendsusetobacco). Note that higher values on FriendsUseTobacco indicate more friends (out of their 5 closest friends) of the participant use tobacco.

```
* Multivariable linear regression of gender and education on

friendsusetobacco

xi: svy: regress friendsusetobacco i.gender i.edu

test _Igender_2 _Iedu_2 _Iedu_3 _Iedu_4 _cons, nosvyadjust

test _Igender_2 _Iedu_2 _Iedu_3 _Iedu_4, nosvyadjust

test _Igender_2, nosvyadjust

test _Iedu_2 _Iedu_3 _Iedu_4, nosvyadjust
```

```
i.gender          _Igender_1-2        (naturally coded; _Igender_1 omitted)
```

i.edu _Iedu_1-4 (naturally coded; _Iedu_1 omitted)

(running regress on estimation sample)

Jackknife replications (50)

----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5

.... 50

Survey: Linear regression

Number of strata = 1

Number of obs = 1,566
Population size = 226,047,519
Replications = 50
Design df = 49
F(4, 46) = 8.26
Prob > F = 0.0000
R-squared = 0.0631

friendsuse~o	Coef.	Jknife * Std. Err.	t	P>|t|	[95% Conf. Interval]	
_Igender_2	0.008756	0.18228	0.05	0.962	-0.35755	0.375062
_Iedu_2	-0.80861	0.352457	-2.29	0.026	-1.5169	-0.10032
_Iedu_3	-0.81445	0.395648	-2.06	0.045	-1.60954	-0.01937
_Iedu_4	-1.44705	0.304695	-4.75	0	-2.05936	-0.83474
_cons	2.196119	0.306155	7.17	0	1.580877	2.81136

Unadjusted Wald test

Unadjusted Wald test

(1) _Igender_2 = 0
(2) _Iedu_2 = 0
(3) _Iedu_3 = 0
(4) _Iedu_4 = 0
(5) _cons = 0

F(5, 49) = 75.43

```
               Prob > F =     0.0000
```

Unadjusted Wald test

```
 Unadjusted Wald test

  (1)  _Igender_2 = 0
  (2)  _Iedu_2 = 0
  (3)  _Iedu_3 = 0

  (4)  _Iedu_4 = 0

      F(  4,    49) =   8.79

         Prob > F =    0.0000
```

Unadjusted Wald test

```
( 1)  _Igender_2 = 0

      F(  1,    49) =    0.00
         Prob > F =    0.9619
```

Unadjusted Wald test

```
  (1)  _Iedu_2 = 0
  (2)  _Iedu_3 = 0

  (3)  _Iedu_4 = 0

       F(  3,    49) =   11.27
         Prob > F =    0.0000
```

From the above table, it can be seen that, compared to those respondents with less than a high school education, those with at least a high school education have a significantly negative linear association with the outcome (i.e., less friends who use tobacco), controlling for all variables in the model. We don’t interpret the gender variable because it is non-significant.

Appendix D: Merging HINTS FDA, Cycle 1 and HINTS FDA, Cycle 2 using SAS

This section provides SAS (Version 9.3 and higher) code for merging the HINTS FDA, Cycle 1 and HINTS FDA, Cycle 2 iterations. It first creates a temporary format for a new ‘survey’ variable that will distinguish between the two iterations. The code then creates two temporary data files and adds the new ‘survey’ variable to each dataset. Next, the two files are merged into one. It will match up variables that have the same name and format and create a merged data file (n = 5,474) that contains one final sample weight (for population point estimates) and 100 replicate weights (NWGT1 TO NWGT100; to compute standard errors).

```
/*FIRST CREATE THE FORMAT FOR THE SURVEY VARIABLE*/
proc format;

value survey
1="FDA-1"
2="FDA-2"
;
run;

/*********************************************************************/

/*CREATE TWO SEPARATE TEMPORARY DATA FILES THAT CONTAIN THE NEW
‘SURVEY’ VARIABLE.

options fmtsearch=(FDA1); /* PUT NAME OF LIBRARY WHERE HINTS FDA CYCLE
                           2 FORMATS ARE STORED*/

data tempFDA1;
set FDA1.hints_fda_09052017_public; /*PUT NAME OF LIBRARY AND NAME OF
                                     EXISTING FDA-1 DATA FILE*/
survey=1;
format survey survey.;
run;

options fmtsearch=(FDA2); /* PUT NAME OF LIBRARY WHERE HINTS FDA CYCLE
2 FORMATS ARE STORED*/

data tempFDA2;
set FDA2.hints_fda2_public; /*PUT NAME OF LIBRARY AND NAME OF
                                          EXISTING FDA-2 DATA FILE*/
survey=2;
format survey survey.;
run;

/**********************************************************************/

/*THIS CODE MERGES THE TWO TEMPORARY DATA SETS CREATED ABOVE. IT ALSO
CREATES ONE FINAL SAMPLE WEIGHT (NWGT0)AND 100 REPLICATE WEIGHTS
(NWGT1 THRU NWGT100)*/

data mergeFDA1_FDA2;
```

```
set tempFDA1 tempFDA2;

/*Create Replicate Weights for trend tests*/

     **Replicate Weights;
     array FDA1Wgts[50] person_finwt1-person_finwt50;
     array FDA2Wgts[50] person_finwt1-person_finwt50;
     array newWghts[100] nwgt1-nwgt100;


**Adjust Final And Replicate Weights;
     if survey eq 1 then do i=1 to 50;  *FDA Cycle 1;
          nwgt0=person_finwt0;
          newWghts[i]=FDA1Wgts[i];
          newWghts[50+i]=person_finwt0;
          end;
     else if survey eq 2 then do i=1 to 50; *FDA Cyle 2;
          nwgt0=person_finwt0;
          newWghts[50+i]=FDA2Wgts[i];
          newWghts[i]=person_finwt0;
          end;
run;


/*********************************************************/

/*YOU CAN USE THE CODE BELOW TO RUN SIMPLE FREQUENCIES ON TWO COMMON
VARIABLES, ‘ECIG_QUIT’ AND ‘ECIG_CHEMICALS’*/


/*SAS CODE*/

proc surveyfreq data=mergefda1_fda2 varmethod=jackknife;
weight nwgt0;
repweights nwgt1-nwgt100 / df = 98 jkcoefs = 0.98;
tables ecig_quit ecig_chemicals;
run;

/*SUDAAN CODE*/

proc crosstab data=mergefda1_fda2 design=jackknife ddf = 98;
weight nwgt0;
jackwgts nwgt1-nwgt100 / adjjack=.98;
class ecig_quit ecig_chemicals;
tables ecig_quit ecig_chemicals;
run;
```

Appendix E: Merging HINTS 5, Cycle 1 and HINTS FDA, Cycle 2 using SAS

This section provides SAS (Version 9.3 and higher) code for merging the HINTS 5, Cycle 1 and HINTS FDA, Cycle 2 iterations. It first creates a temporary format for a new ‘survey’ variable that will distinguish between the two iterations. The code then creates two temporary data files and adds the new‘survey’ variable to each dataset. Next, the two files are merged into one. It will match up variables that have the same name and format and create a merged data file (n = 5,021) that contains one final sample weight (for population point estimates) and 100 replicate weights (NWGT1 TO NWGT100; to compute standard errors).

```
/*FIRST CREATE THE FORMAT FOR THE SURVEY VARIABLE*/
proc format;

value survey
1="HINTS 5 CYCLE 1"
2="FDA-2"
;
run;

/*********************************************************************/

/*CREATE TWO SEPARATE TEMPORARY DATA FILES THAT CONTAIN THE NEW
‘SURVEY’ VARIABLE.*/

options fmtsearch=(HINTS5C1); /*PUT NAME OF LIBRARY WHERE HINTS 5
                                 CYCLE 1 FORMATS ARE STORED*/
data tempHINTS5CYCLE1;
set HINTS5C1.hints5cycle1_public; /*PUT NAME OF LIBRARY AND NAME OF
                                       EXISTING HINTS 5 CYCLE 1 FILE*/
survey=1;
format survey survey.;
run;

options fmtsearch=(FDA2); /* PUT NAME OF LIBRARY WHERE HINTS 5 CYCLE 1
FORMATS ARE STORED*/

data tempFDA2;
set FDA2.hints_fda2_public; /*PUT NAME OF LIBRARY AND NAME
                                       OF EXISTING FDA-2 DATA FILE*/
survey=2;
format survey survey.;
run;

/**********************************************************************/

/*THIS CODE MERGES THE TWO TEMPORARY DATA SETS CREATED ABOVE. IT ALSO
CREATES ONE FINAL SAMPLE WEIGHT (NWGT0) AND 100 REPLICATE WEIGHTS
(NWGT1 THRU NWGT100)*/

data mergeHINTS5C1_FDA2;
set tempHINTS5CYCLE1 tempFDA2;
```

```
/*Create Replicate Weights for trend tests*/

      **Replicate Weights;
      array hints5wgts[50] person_finwt1-person_finwt50;
      array fda2wgts[50] person_finwt1-person_finwt50;
      array newWghts[100] nwgt1-nwgt100;


**Adjust Final And Replicate Weights;
      if survey eq 1 then do i=1 to 50;  *HINTS 5 CYCLE 1;
            nwgt0=person_finwt0;
            newWghts[i]=hints5wgts[i];
            newWghts[50+i]=person_finwt0;
            end;
      else if survey eq 2 then do i=1 to 50; *FDA-2;
            nwgt0=person_finwt0;
            newWghts[50+i]=fda2wgts[i];
            newWghts[i]=person_finwt0;
            end;
run;


/********************************************************/

/*YOU CAN USE THE CODE BELOW TO RUN SIMPLE FREQUENCIES ON TWO COMMON
VARIABLES, ‘LOTOFEFFORT’ AND ‘TRUSTDOCTOR’*/


/*SAS CODE*/

proc surveyfreq data=mergeHINTS5C1_FDA2 varmethod=jackknife;
weight nwgt0;
repweights nwgt1-nwgt100 / df = 98 jkcoefs = 0.98;
tables lotofeffort trustdoctor;
run;

/*SUDAAN CODE*/

proc crosstab data=mergeHINTS5C1_FDA2 design=jackknife ddf = 98;
weight nwgt0;
jackwgts nwgt1-nwgt100 / adjjack=.98;
class ecig_quit ecig_chemicals;
tables lotofeffort trustdoctor;
run;
```