Software for Analysis of YRBS Data

September 2024

Where can I get more information? Visit www.cdc.gov/yrbss or call 800–CDC–INFO (800–232–4636).

CONTENTS

Overview

Youth Risk Behavior Surveys (YRBS) employ a complex sampling design. Therefore, to analyze YRBS data correctly, statistical software packages that account for this sampling design must be used. This document describes six selected statistical software packages appropriate for analyzing YRBS data: SUDAAN, SAS, Stata, SPSS, Epi Info, and R. For each statistical software package, information on analytic capabilities, data requirements, variance estimation, and survey degrees of freedom is provided along with sample design statements and a sample program. Tables 1 and 2 provide a comparison of features across the selected five statistical software packages. Table 3 compares the results of national YRBS analyses using procedures within each statistical software package.

This document is intended for analysts familiar with statistical software packages and with YRBS data in general. It does not explain all details and issues related to analyzing YRBS data or how to use all procedures available in each statistical software package. It does not include information on all versions of these software packages; however, it is assumed that later versions of each package will have at least the same capabilities as previous versions. For that reason, software packages are described as a version number "and higher."

Background

Analysis of data from surveys that employ a complex sampling design, such as the YRBS, must account for the sampling design (stratification, clustering, and unequal selection probabilities) to obtain valid point estimates, standard errors, confidence intervals, and tests of hypotheses. Simply doing a weighted analysis using statistical software programs like SAS Proc Means or Proc Freq is not appropriate because the variance estimation and hypothesis testing in such programs use formulas appropriate for simple random sampling. These formulas do not account for unequal sampling weights (unequal probabilities of selection), stratification, and clustering. Even if standardized weights, which are scaled to total to the sample size rather than the population size as in the national YRBS, are used, the variance estimation and hypothesis tests are still not valid. Variance may be either underestimated (which usually occurs when sampling designs include clustering and unequal probabilities of selection) or overestimated (which can occur with stratification in an unclustered sampling design).

Several statistical software packages are designed to analyze complex sample survey data correctly. SUDAAN from Research Triangle Institute, WesVar from Westat Incorporated, and IVEware from the University of Michigan Survey Research Center are three such statistical software packages that are designed specifically for analysis of complex sample survey data. General use statistical software packages -- including SAS, Stata, SPSS, Epi Info, and R – also have developed special procedures or modules to analyze complex sample survey data.

For general information on analysis of complex sample survey data, refer to Section E, Chapter 19 of the United Nations book – *Household Sample Surveys in Developing and Transition Countries*, available at: http://unstats.un.org/unsd/HHsurveys/ or the other resources listed at the end of this document. For additional information on YRBS data and methodology, refer to the CDC's YRBS website at http://www.cdc.gov/yrbss.

1. SUDAAN

SUDAAN is designed to analyze data from complex surveys and experimental studies. SUDAAN version 11 and higher offers analysis capabilities that include cross-tabulation, frequency, ratio, and multiple regression modeling techniques. SUDAAN, like SAS, requires that syntax be written; no graphical user interface is available to allow menu-driven (i.e., point-and-click) analysis.

Note: SUDAAN is available in stand-alone and SAS-callable versions. SAS-callable SUDAAN is run by including SUDAAN statements in a SAS program. This is convenient when working in SAS for data management, since the user does not have to exit SAS and open SUDAAN to run analyses.

1.1. Analytic capabilities: SUDAAN has a wide range of analytic capabilities. Descriptive analyses include means, geometric means, medians and other percentiles, totals, ratios, and proportions. All of these produce standard errors and confidence intervals. Asymmetric confidence intervals are produced for proportions using either Proc Crosstab, Proc Descript, or Proc Vargen (Proc Vargen is available in version 11 and higher). Standardized means and rates also can be obtained. Estimates for domains are obtained by using a TABLES statement that includes one or more categorical variables. Domain estimates can be compared via system or user-defined linear contrasts. Crosstabulations include odds ratios, relative risks, chi-square tests (Pearson type and log-linear), Cohen's Kappa measure of agreement, and the Cochran-Mantel-Haenszel tests for single and stratified two-way tables. Regression analyses available include general linear models, binary and polychotomous logistic regression (both ordinal and nominal), survival analysis, and log-linear models. The SUBPOPN statement can be used with any procedure to obtain estimates for a subpopulation. SUDAAN has an extensive capability to estimate and test user-specified contrast matrices on population parameters, including regression coefficients. It also has procedures for analyzing multiply imputed datasets, so that the variance due to multiple imputation can be included in the variance estimate. Design effect can be obtained for a variety of estimated statistics.

1.2. Data requirements: All variables used in analyses, including the sample design variables (stratum, primary sampling unit (PSU), and weight variables), must be numeric; character variables are not recognized even if their values are numbers. Input data files can be SAS, SPSS, or ASCII. Data should be sorted by the variables that appear on the NEST statement (stratum and PSU variables) before analysis, otherwise procedure syntax must contain the NOTSORTED option when specifying input data sets. All independent variables must be coded >0, e.g., a binary variable should be coded (1,2) rather than (0,1).

1.3. Variance estimation: Variance estimation options available in SUDAAN are Taylor Series Linearization (TSL) and two replication methods, balanced repeated replication and jackknife; the default is TSL. A finite population correction can be included at any stage of sampling for without replacement sampling designs. If an analysis includes data from one or more strata that contain only a single PSU, the analysis will not proceed and a warning will appear in the log. For such analyses the MISSUNIT option can be added

to the NEST statement and SUDAAN will obtain the variance contribution for such units using the difference between that unit's value and the overall mean value of the population. The only other option for variance estimation in such a situation is for the user to collapse strata to eliminate strata with only one PSU.

1.4. Survey degrees of freedom: SUDAAN defines survey degrees of freedom as the number of PSUs minus the number of first stage sampling strata. Thus, when data on an analysis variable are missing for all sampled elements in one or more PSU or stratum, which most commonly occurs when analyses are performed for a small subpopulation, the degrees of freedom will be overestimated. The overestimation can be remedied by using the atlevel1 and atlevel2 options on the Proc statement to determine the number of strata and PSUs included in an analysis and rerunning the analysis with the correct number of degrees of freedom indicated to SUDAAN using the DDF= option on the PROC statement. Using the correct number of survey degrees of freedom is important because this statistic is used to determine the critical value from the t distribution that will be used to construct confidence intervals. If the survey degrees of freedom are overestimated, a smaller critical value than appropriate will be used to calculate confidence intervals, resulting in confidence intervals that are narrower than they should be.

1.5. Sampling designs: Multiple design options allow data from stratified, clustered, or multistage sampling designs to be analyzed. Sample members may have been selected with unequal probabilities and either with or without replacement. Any number of strata and sampling stages can be specified. In addition, different design options may be combined in one study if different sampling methods were used for different parts of the population. The user describes the sample survey design in three statements: (1) by specifying an option for the DESIGN keyword on the PROC statement, (2) by specifying the stratification and clustering (PSU) variables on the NEST design statement, and (3) by specifying the analysis weight variable on the WEIGHT design statement. The default design option is WR (with replacement at first stage), which is appropriate for analysis of YRBS data and many other national and state survey data sets that use multistage sampling designs, such as the Behavioral Risk Factor Surveillance System (BRFSS), the National Health and Nutrition Examination Survey (NHANES), and the National Health Interview Survey (NHIS). The sample design statements must be included in the syntax each time an analysis is run.

```
PROC …… design = [WR|WOR|UNEQWOR|STRWR|STRWOR|SRS|BRR|JACKKNIFE];
NEST stratification_variable  PSU_variable;
WEIGHT analysis weight_variable;
```

1.6. Sample program code: Program code used for the analyses that appear in Table 3 is provided below. Data must be sorted by the stratification variable and cluster/PSU. If data is not sorted, the NOTSORTED option must be included in the syntax each time an analysis is run.

```
proc crosstab data=yrbs23 design=wr NOTSORTED;
nest stratum psu / missunit ;
weight weight ;
class qn8 qn56 qn52 / nofreqs ;
tables qn8 qn56 qn52 ;
print  / style=NCHS rowperfmt=F9.4 serowfmt=F9.4 uprowfmt=F9.4
      lowrowfmt=F9.4 ;
run;
```

2. SAS

SAS versions 8 and higher include special sample survey procedures that are appropriate for analyzing complex survey data like the YRBS. These sample survey procedures use SAS syntax that will be familiar to those who are already SAS users. SAS, like SUDAAN, requires that syntax be written; no graphical user interface is available to allow menu-driven (i.e., point-and-click) analysis.

2.1. Analytic capabilities: SAS (version 9.4 and higher) sample survey analysis capabilities include descriptive statistics (means, ratios, totals, and proportions with standard errors and confidence intervals, population quantiles), crosstabulations for 2-way and n-way tables with measures of relative risks and tests of independence (Wald test, Rao-Scott likelihood ratio test, and Rao-Scott chi-square test), generalized linear regression, logistic regression, and survival analysis. Design effect can be calculated for the proportion estimate and the regression coefficient estimates. The following regression models are available in Proc SurveyLogistic: binary logistic regression and ordered and nominal polychotomous logistic regression. Proc SurveyMeans does not include a 2-sample t-test for domain comparisons; however, these can be obtained using Proc SurveyReg. Proc SURVEYMEANS also estimates percentiles, with the variance of percentiles being estimated using Woodruff methods (Dorfman and Valliant 1993; Särndal, Swensson, and Wretman 1992; Francisco and Fuller 1991). Symmetric confidence intervals are produced for proportions. The DOMAIN statement with one or more categorical variables is used to obtain estimates for domains in all procedures except procedure SURVEYFREQ, for which domain analysis can be obtained by cross-tabulating the domain variable with the analysis variables. SAS does not have a statement that allows a subpopulation (e.g., 9th grade female students) to be analyzed, however, subpopulation analyses can be performed by first creating an indicator variable (e.g., NINTHFEM) that indicates whether a sample element belongs to the subpopulation. Then the statement DOMAIN NINTHFEM can be used to obtain the desired analysis. Domain and subpopulation analyses should not be attempted using the BY, IF, or WHERE statements because this will result in inappropriate subsetting of the data. Variance estimates, confidence intervals, and tests of hypothesis from such analyses are invalid.

2.2. Data requirements: Not all variables used in analyses must be numeric. Categorical variables can be either numeric or character, only continuous variables must be numeric. SAS data files are used for analysis (.sas7bdat). The input data file does not need to be sorted by stratum and/or primary sampling unit (PSU) variables before analysis.

2.3. Variance estimation: Variance estimation options available in SAS are Taylor Series Linearization (TSL) and two replication methods, balanced repeated replication (BRR) and jackknife; the default is TSL. A finite population correction term can be applied for single stage sampling designs such as stratified random sampling and simple random sampling. If an analysis includes data from one or more strata that contain only a single PSU, the analysis will proceed and a note will appear in the log. The note indicates that one or more strata contained a single PSU and that single-PSU strata are not included in

the variance estimates. The only other option for variance estimation in such a situation is for the user to collapse strata to eliminate strata with only one PSU. The exception is procedure SURVEYREG. To estimate stratum variances, the procedure, by default, collapses or combines those strata that contain only one PSU. If you specify the NOCOLLAPSE option in the STRATA statement, PROC SURVEYREG does not collapse strata and uses a variance estimate of zero for any stratum that contains only one PSU.

2.4. Survey degrees of freedom: SAS defines survey degrees of freedom as the number of PSUs minus the number of first stage sampling strata among strata and PSUs that contain at least one observation with a value for the analysis variable(s), an alternate definition recommended by Korn and Graubard (1999) in the context of subpopulation analysis. Thus, when data on an analytic variable are missing for all respondents in one or more PSU or stratum, which most commonly occurs when performing analyses for a small subpopulation, the degrees of freedom will be calculated correctly by SAS, not overestimated, and there is no need to apply a remedy as per SUDAAN. A note in the log indicates that there were empty clusters for a variable and how many clusters were included in the analysis.

2.5. Sampling designs: There are three sample design statements in SAS where the information captured on the NEST and WEIGHT statements in SUDAAN is entered: CLUSTER, where the name of the PSU variable is placed; STRATA, where the name of the stratification variable(s) is placed; and WEIGHT, where the name of the analysis weight variable is placed. Information on clustering and stratification can be entered for only the first stage of sampling. For complex samples, SAS sample survey procedures assume a with-replacement sampling design, which is the equivalent of specifying DESIGN = WR in SUDAAN. This sampling design is appropriate for analysis of YRBS and many other national and state data sets that use multistage sampling designs, such as the Behavioral Risk Factor Surveillance System (BRFSS), the National Health and Nutrition Examination Survey (NHANES), and the National Health Interview Survey (NHIS). Less complex sampling designs can be described to SAS by omitting specific design statements. For example, a single-stage stratified random sampling design can be indicated by omitting the CLUSTER statement and a design with no stratification at the first stage can be indicated by omitting the STRATA statement. An unweighted design can be indicated by omitting the WEIGHT statement and simple random sampling can be indicated by omitting all three sample design statements. The appropriate sample design statements (if any) must be included in the syntax each time an analysis is run.

```
STRATA stratification variable;
CLUSTER PSU variable;
WEIGHT analysis weight variable;
```

2.6. Sample program code: Program code used for the analyses that appear in Table 3 is provided below. The log indicates if there are any empty clusters omitted from the analysis.

```
proc surveyfreq data=yrbs23;
strata stratum ;
cluster psu ;
weight weight ;
tables qn8 qn56 qn52 / cl;
run ;
```

3. Stata

Stata (version 7.0 and higher) offers the capability to perform many statistical procedures on complex sample survey data, and graphics capabilities as well. Stata, like SUDAAN and SAS, can be run using syntax, but a graphical user interface (GUI) is available that also allows analysis to be menu driven (i.e., point-and-click).

3.1. Analytic capabilities: Stata offers a wide range of analyses for sample survey data, with mathematical statistical capabilities for user-specified contrast matrices on population parameters including regression coefficients. Thus it possesses analytic capabilities similar to those available in SUDAAN and offers some regression models that are not available in SUDAAN. Design effect can be obtained for a variety of estimated statistics. Descriptive statistics (means, ratios, totals, and proportions) with standard errors and confidence intervals and crosstabulations with Rao-Scott corrected chi-square test are available. In addition, a number of regression analyses are available including linear regression; generalized linear regression; tobit and probit models; Poisson, negative binomial, and zero-inflated Poisson models; binary and polychotomous (both ordered and nominal) logistic regression; structural equation and multilevel modeling, and survival analysis. Domain estimates can be obtained using OVER on the command line and subpopulation analyses can be performed using the SUBPOP option. The Bonferroni multiple comparisons procedure is also available for hypothesis testing with survey data. In version 13.0 and higher, Stata produces asymmetric confidence intervals for proportions and tabulations using a logit transform, for example by using svy: tabulate (Statistics… Survey data analysis… Tables… One-way tables). Stata version 12 and earlier produced symmetric confidence intervals for proportions using svy: proportion (Statistics… Survey data analysis… Means, proportions, ratios, totals … Proportions). Stata also includes an imputation option, which allows missing data to be filled in using regression models and procedures for analyzing multiply imputed datasets, so that the variance due to multiple imputation can be included in the variance estimate.

3.2. Data requirements: Although variables included in Stata data sets can be either numeric or character, all variables used in an analysis must be numeric. Stata (.dta) data files are used for analysis. CDC does not provide YRBS data as Stata data files but ACCESS data files are available. Data from an ACCESS table can be imported using the ODBC command and saved as a Stata (.dta) file. Data from an ACCESS table may also be converted to an EXCEL file and data from the EXCEL file can then be copied into the Stata data editor and saved as a Stata .dta file. SAS files can be easily converted to Stata .dta files using the Stat/Transfer or DBMSCOPY software packages. Stata will also read in SAS transport files. The input data file does not need to be sorted by stratum and/or primary sampling unit (PSU) variables before analysis.

3.3. Variance estimation: Variance estimation options available in Stata are Taylor Series Linearization (TSL) and two replication methods, balanced repeated replication (BRR) and jackknife; the default is TSL. Version 12.0 and higher also includes the options bootstrap (with replicate weights) and successive difference replication. A finite population correction can be included for random sampling without replacement of

sampling units within strata. When declaring the survey design for a data set, the user also can indicate how variance estimation should be handled for a single-PSU stratum. The options are: report missing standard errors (default), treat the sampling unit as certainty units, scale variance using certainty units, and center using grand mean (equivalent to what SUDAAN does). If none of these options is acceptable for variance estimation, the user can collapse strata to eliminate strata with only one PSU. The default is to report missing standard errors. Choice of variance estimation method (TSL, BRR, JACKKNIFE) also can be indicated when declaring the survey design.

3.4. Survey degrees of freedom: Stata defines survey degrees of freedom as the number of PSUs minus the number of first stage sampling strata among strata and PSUs that contain at least one observation with a value for the analysis variable(s), an alternate definition recommended by Korn and Graubard (1999) in the context of subpopulation analysis. Thus, when data on an analysis variable are missing for all respondents in one or more PSU or stratum, which most commonly occurs when performing analyses for a small subpopulation, the degrees of freedom will be calculated correctly by Stata, not overestimated, and there is no need to apply a remedy as per SUDAAN.

3.5. Sampling designs: Stata allows a variety of complex sampling designs including multistage, stratified, and clustered sampling with and without replacement. Any number of strata and sampling stages can be specified. When performing menu-driven analyses, the information captured on the NEST and WEIGHT statements in SUDAAN is entered into boxes on the MAIN (PSU and stratification variables) and WEIGHT (analysis weight variable) tabs of the dialogue box that appears after "Declare survey design for data set" is chosen from the Survey Data Analysis menu. If syntax is written the information is included on the SVYSET statement. Once this information has been entered or the SVYSET statement has been run for a data set, it does not need to be re-entered or included in the syntax for each analysis during that Stata session as long as that data set is open.

```
svyset [pweight=analysis weight variable], strata(stratification
variable)  psu(PSU variable)
```

3.6. Sample program code: Although analyses can be run using a GUI in Stata, the corresponding syntax code also is available. The program code used for the analyses that appear in Table 3 is provided below.

```
use "K:\Stata Data\YRBS2023.dta", clear
svyset psu [pweight=weight], strata(stratum) vce(linearized) singleunit(centered)
svy linearized : proportion qn8
svy linearized : proportion qn56
svy linearized : proportion qn52
```

This code was generated by the following actions:

- Open Stata
- From the **File** menu

  - Select **Open**
  - Locate the YRBS 2023 data file and click **Open**
- From the **Statistics** menu select **Survey data analysis… Setup and utilities … Declare survey design for dataset**
  - In the **Svyset – Survey data setting** dialogue box that appears
    - On the **Main** tab
      - Select 1 from the drop-down list in the **Number of stages** box
      - For **Stage 1:**
        - Select psu from the drop-down list in the **Sampling units** box
        - Select stratum from the drop-down list in the **Strata** box
    - On the **Weights** tab
      - Select **Sampling weight variable** under **Weight type**
      - Select weight from the drop-down list in the **Sampling weight variable** box
    - On the **SE** tab
      - Highlight **Linearized** in the **Method for variance estimation** box
      - Under **Stratum with a single sampling unit** select **Center at the grand mean**
    - Click OK

A listing of the information entered will appear in the **Results** window

- From the **Statistics** menu select **Survey data analysis… Means, proportions, ratios, totals… Proportions**
  - In the **svy: proportions – Estimate proportions for survey data** dialogue box
    - On the **SE/Cluster** tab
      - Verify that the **Survey data estimation** box is checked
      - Verify that **Linearized** is highlighted in the **Standard error type** box
    - On the **Reporting** tab verify that 95 appears in the **Confidence level** box
    - On the **Model** tab
      - Select QN8 from the drop-down list in the **Variables** box and click **Submit**
      - After output for QN8 appears in the **Results** window, replace QN8 with QN56 in the **Variables** box and click **Submit**
      - After output for QN56 analysis appears in the **Results** window, replace QN56 with QN52 in the **Variables** box and click **Submit**

The corresponding syntax which appears in the **Review** window (upper left of screen) under **Command** can be saved to a file for later use. To do this, the user must open a log file at the beginning of the session (**File… Log… Begin…** then choose a location and file name for your log).

Note that the syntax above results in asymmetric confidence intervals in Stata version 13 and later, but produces symmetric confidence intervals in Stata versions 12 and earlier. To obtain asymmetric confidence intervals for proportions in all versions of Stata, one can also use **svy: tabulate (Statistics… Survey data analysis… Tables… One-way tables)**. The corresponding syntax is:

```
use "K:\Stata Data\YRBS2023.dta", clear
svyset psu [pweight=weight], strata(stratum) vce(linearized) singleunit(centered)
svy linearized : tabulate qn8, cell se ci obs
svy linearized : tabulate qn56, cell se ci obs
svy linearized : tabulate qn52, cell se ci obs
```

Cautionary note: When obtaining frequencies for a list of two or more variables, Stata uses list-wise exclusion and generates estimates using responses only from observations with no missing data for any of the variables on the list. This will be readily apparent from the unweighted sample sizes, which are all the same. To obtain proportions that use all non-missing data for each variable (i.e., pair-wise or table-by-table exclusion) the analysis must be run separately for each variable as shown above.

4. SPSS

SPSS has an add-on module, SPSS Complex Samples, which includes sample selection and analysis of complex sample survey data. The version (21) covered in this review has analysis capabilities that are somewhat more limited than those of SUDAAN or Stata. Like STATA and Epi Info, SPSS can be run using syntax but also includes a graphical user interface (GUI) that allows analysis to be menu driven (i.e., point-and-click).

4.1. Analytic capabilities: SPSS offers a number of analyses for sample survey data, but fewer than those currently available in SUDAAN and Stata. Descriptive statistics (means, ratios, totals, and proportions) with standard errors and confidence intervals and 2-way crosstabulations (with odds ratios, relative risks, and tests of independence) are available. Subpopulation analysis is available. In addition, regression analyses including general linear models (with analysis of variance and analysis of covariance models); binary, ordinal polychotomous, and multinomial logistic regression models; and survival analysis using Cox regression are available. Design effect can be calculated for estimated statistics. Asymmetric confidence intervals are produced for proportions. Several multiple comparisons procedures are also available in the general linear models capability for hypothesis testing with survey data.

4.2. Data requirements: Not all variables used in analyses must be numeric. Categorical variables can be either numeric or character, continuous variables must be numeric. SPSS (.sav) data files are used for analysis. SPSS data files for the national YRBS are available from CDC. SPSS data files for the state and local YRBS are not available from CDC, but syntax to produce SPSS data files from ASCII data is available. The input data file does not need to be sorted by stratum and/or primary sampling unit (PSU) variables before analysis.

4.3. Variance estimation: Only Taylor Series Linearization is available for variance estimation in SPSS. A finite population correction can be included as part of the sampling plan for without replacement sampling designs. If an analysis includes data from one or more strata that contain only a single PSU, the analysis will proceed and no warning will appear in the log; variance will be estimated with no contribution from the singleton PSUs. The only other option for variance estimation with singleton PSUs is for the user to collapse strata to eliminate strata with only one PSU.

4.4. Survey degrees of freedom: SPSS defines survey degrees of freedom as the number of PSUs minus the number of first stage sampling strata among strata and PSUs that contain at least one observation with a value for the analysis variable(s), an alternate definition recommended by Korn and Graubard (1999) in the context of subpopulation analysis. Thus, when data on a variable are missing for all respondents in one or more PSU or stratum, which most commonly occurs when performing analyses for a small subpopulation, the degrees of freedom will be calculated correctly by SPSS and there is no need to apply a remedy as per SUDAAN.

4.5. Sampling designs: SPSS accommodates a variety of complex sampling designs including multistage, stratified, and clustered sampling with and without replacement. For multistage designs, information on up to three stages of sampling can be included, with stratification and clustering at each stage. Unequal-probability without replacement (UnequalWOR) designs including 2-PSU-per-stratum designs are also supported (for example, via Brewer's Method). When performing menu-driven analysis, the information captured on the NEST and WEIGHT statements in SUDAAN is entered into boxes in a dialogue box when preparing for analysis and is saved as a sampling plan for the data set. A sampling design for variance estimation - WR, EqualWOR, or UnequalWOR - also is specified as part of the sampling plan. Once the sampling plan for a data set has been created, it will be opened along with the data set at the beginning of an SPSS session and the information will be retained for all analyses performed during that session or until another data set is opened during the session.

4.6. Sample program code: Although analyses can be run using a GUI in SPSS, the corresponding syntax code also can be pasted to a program file and run. The program code used for the analyses that appear in Table 3 is provided below.

```
 * Analysis Preparation Wizard.
CSPLAN ANALYSIS
  /PLAN FILE='C:\Users\xgj4\Desktop\YRBS2023plan.csaplan'
  /PLANVARS ANALYSISWEIGHT=weight
  /SRSESTIMATOR TYPE=WOR
  /PRINT PLAN
  /DESIGN STRATA=stratum CLUSTER=psu
  /ESTIMATOR TYPE=WR.


* Complex Samples Frequencies.
CSTABULATE
 /PLAN FILE='K:\SPSS data\YRBS2023plan.csaplan'
 /TABLES VARIABLES= QN8 QN56 QN52
 /CELLS POPSIZE TABLEPCT
 /STATISTICS SE CIN(95) COUNT
 /MISSING SCOPE=TABLE CLASSMISSING=EXCLUDE.
```

This code was generated by the following actions:

- Open SPSS
- From the **File** menu, select **Open**… **Data,** locate and select the YRBS 2023 data file (yrbs23.sav) and click **Open**
- From the **Analyze** menu select **Complex Samples… Prepare for Analysis**
  - In the **Analysis Preparation Wizard** dialogue box select **Create a plan file** and click **Browse**
    - In the **Save As** dialogue box enter YRBS2023plan in the **File Name** box and click **Save**
  - In the **Analysis Preparation Wizard** dialogue box click **Next**

  - In the **Analysis Preparation Wizard** dialogue box under **Stage 1: Design Variables**
    - Highlight *Stratum [stratum]* on the list in the **Variables** box and click the arrow next to the **Strata** box to move the stratification variable into that box
    - Highlight *Primary Sampling Unit [psu]* on the list in the **Variables** box and click the arrow next to the **Clusters** box to move the PSU variable into that box
    - Highlight *Weight [weight]* on the list in the **Variables** box and click the arrow next to the **Sample Weight** box to move the sampling weight variable into that box
    - Click **Next**
  - In the **Analysis Preparation Wizard** dialogue box under **Stage 1: Estimation Method**, select **WR** and click **Next**
  - In the **Analysis Preparation Wizard** dialogue box under **Stage 1: Plan Summary** verify that the correct information has been entered and click **Next**
  - Under **Completing the Analysis Wizard** in the **Analysis Preparation Wizard** dialogue box, select **Save your specifications to a plan file** and click **Finish**

A summary of the plan will appear in the **Output** window at this point

- Minimize the **Output** window
- From the **Analyze** menu select **Complex Samples… Frequencies**
  - When the **Complex Samples Plan for Frequencies Analysis** dialogue box appears, verify that the name of the sampling plan file appears in the **File** box under **Plan**, then click **Continue**
    - If the sampling plan file name does not appear, click **Browse…** , locate and select the sampling plan file, then click **Continue**
  - In the **Complex Samples Frequencies** dialogue box
    - Highlight *Did not always wear a seat belt [QN8]* on the list in the **Variables** box and click the arrow next to the **Frequency Tables** box to move the analysis variable into that box
    - Repeat for *Ever had sexual intercourse [QN56]* and *Used heroin 1+ times in life [QN52]*
    - Click on **Statistics**
      - Under **Cells** check the boxes for **Population size** and **Table percent**
      - Under **Statistics**
        - Check the boxes for **Standard error** and **Unweighted count**
        - Check the box for **Confidence interval** and verify that 95 appears in the **Level** box, then click **Continue**

- Click on **Missing Values**
  - Under **Tables** verify that **Use all available data (table-by-table deletion)** is selected
  - Under **Categorical Design Variables** verify that **User-missing values are invalid** is selected, then click **Continue**
- Click **OK** in the **Complex Samples Frequencies** dialogue box to run the analysis

Results will appear in the **Output** window. Clicking **Paste** in the previous step instead of **OK** causes the syntax to appear in a **Syntax** window to be run and saved for later use.

5. Epi Info

Epi Info includes a module for complex sample survey analysis. The analytic capabilities are quite limited and specifically oriented towards public health field work applications. Like Stata and SPSS, Epi Info can be run using syntax but includes a graphical user interface (GUI) that also allows analysis to be menu driven (i.e., point-and-click).

5.1. Analytic capabilities: Epi Info (version 7.2) sample survey analysis capabilities are limited to means, proportions, 2-way crosstabulations, odds ratios, risk ratios, and risk differences, with standard errors and confidence intervals. These estimates also are provided for domains formed by levels of a categorical variable. Epi Info will also estimate the difference between domain means, with an estimated standard error for the difference and a confidence interval on the population difference. Design effect can be calculated for the proportion estimate. Epi Info does not include a test of independence for crosstabulations and does not estimate population totals.

5.2. Data requirements: Not all variables used in analyses must be numeric. Categorical variables can be either numeric or character, continuous variables including the weight variable must be numeric. ACCESS data files are used for analysis. The input data file does not need to be sorted by stratum and/or primary sampling unit (PSU) variables before analysis.

5.3. Variance estimation: Only Taylor Series Linearization is available for variance estimation. No finite population correction is calculated, so either sampling fractions must be small or sampling must be with replacement. If an analysis includes data from one or more strata that contain only a single PSU, the analysis will proceed. This situation is handled by not including single-PSU strata in the variance estimates. The only other option for variance estimation in such a situation is for the user to collapse strata in order to eliminate strata with only one PSU.

5.4. Survey degrees of freedom: Epi Info defines survey degrees of freedom as the number of PSUs minus the number of first stage sampling strata among strata and PSUs that contain at least one observation with a value for the analysis variable(s), an alternate definition recommended by Korn and Graubard (1999) in the context of subpopulation analysis. Thus, when data on a variable are missing for all respondents in one or more PSU or stratum, which most commonly occurs when performing analyses for a small subpopulation, the degrees of freedom will be calculated correctly by Epi Info and there is no need to apply a remedy as per SUDAAN.

5.5. Sampling designs: Epi Info accommodates sampling designs including stratified sampling with or without clustering, multistage samples, and unequal-probability (e.g., probability proportional to size) samples. Sample design information captured on the NEST and WEIGHT statements in SUDAAN is entered into the appropriate box (Weight, PSU, Stratify by) in the dialog box that appears once an analysis (Complex Sample Frequencies, Complex Sample Tables, Complex Sample Means) has been selected. Information on clustering and stratification can be entered for only the first

stage of sampling. Epi Info complex sample procedures assume a common sampling design which is the equivalent of specifying DESIGN = WR in SUDAAN. This sampling design is appropriate for analysis of many national and state data sets in addition to YRBS such as the Behavioral Risk Factor Surveillance System (BRFSS), the National Health and Nutrition Examination Survey (NHANES), and the National Health Interview Survey (NHIS) which use multistage sampling designs. The sample design variables must be entered in the dialog box or included in the syntax each time an analysis is run.

5.6. Sample program code: Although analyses can be run using a GUI in Epi Info, the corresponding syntax code also is available. The program code used for the analyses that appear in Table 3 is provided below.

```
READ 'C:\Epi_Info_Data\yrbs23.mdb':XXHqn
FREQ  qn8 qn56 qn52 STRATAVAR=stratum WEIGHTVAR=weight PSUVAR=psu
      OUTTABLE=results
```

This code was generated by the following actions:

- Open Epi Info and click **Analyze Data… Classic**
- In the **Analysis** window on the left side of the screen, under **Analysis Commands… Data**, click **Read (import)**
  - In the **READ** dialogue box, select the **Database Type**, then locate the YRBS 2023 data file and click **OPEN**
  - In the next **READ** dialogue box (contains the name of the data file), highlight **XXHqn** (name of the table that contains the YRBS data) on the list of tables under **All**
  - Click **OK**
- In the **Analysis** window under **Analysis Commands… Advanced Statistics**, click **Complex Sample Frequencies**
  - In the **Complex Sample Frequencies** dialogue box
    - Select weight from the drop-down list in the **Weight** box
    - Select psu from the drop-down list in the **PSU** box
    - Select QN8, QN56 and QN52 from the drop-down list in the **Frequency of** box
    - Select stratum from the drop-down list in the **Stratify by** box
    - Fill out **Output to Table** with a table name to save the analyses, such as “results”
    - Click **OK**

The analysis will run and the results appear in the **Analysis Output** window at the top of the screen. The corresponding syntax appears in the **Program Editor** window below the **Output** window and can be saved to a file for later use.

To see the standard error and confidence limits of the complex sample frequencies, open the source file ‘yrbs23.mdb’ and the new table “results” that was added to it using the **Output to Table** option.

6. R

Many thanks to Dr. Thomas Lumley for developing and contributing the original version of this extension, which has been lightly edited/updated.

R is a free software that can accommodate multiple statistical packages (http://www.r-project.org, R Foundation 2009). To analyze YRBS data, it is necessary to download a complex survey analysis package. This chapter covers the "survey" package in R, version 4.0 (http://r-survey.r-forge.r-project.org/survey/index.html, Lumley 2004, 2009, requires R 3.1.0 or higher), which offers the ability to perform many statistical procedures on complex sample survey data. R is well known for its graphics capabilities as well.

6.1. Analytic capabilities: the R survey package offers a wide range of analyses for sample survey data, with mathematical statistical capabilities for user-specified contrast matrices on population parameters including regression coefficients. Thus it possesses analytic capabilities similar to those available in SUDAAN and offers some regression models that are not available in SUDAAN. Design effect can be obtained for a variety of estimated statistics. Descriptive statistics (means, ratios, totals, quantiles, and proportions) with standard errors and confidence intervals and crosstabulations with Rao-Scott corrected chi-square test are available. In addition, a number of regression analyses are available including linear regression, generalized linear regression, probit models, Poisson models, binary and ordered logistic regression, loglinear models, and survival analysis. The R survey package can also perform factor analysis and principal components analysis. The svyciprop() function can be used to calculate asymmetric confidence intervals for proportions in complex survey designs, with a method specification "xlogit". The survey package svyciprop() has facilities for analyzing multiply-imputed data, and other R packages such as 'mice' and 'mi' assist in creating these multiple imputations.

6.2. Data requirements: Variables used in analysis can be numeric, character, or factor (representing categorical variables). R can read in Stata .dta files, SAS transport files, SPSS .sav files, and (under Windows) can directly query Access data tables. The input data file does not need to be sorted by stratum and/or primary sampling unit (PSU) variables before analysis. Using the library "haven" or "foreign" will allow you to download any data formats.

6.3. Variance estimation: Variance estimation options available in R are Taylor Series Linearization (TSL); replication methods including balanced repeated replication (BRR), jackknife, and bootstrap; or user-supplied. The variance estimation method is specified when the survey design is described. A finite population correction can be included for random sampling without replacement of sampling units within strata, or for PPS sampling without replacement. A global option controls how variance estimation should be handled for a single-PSU stratum. The options are: report missing standard errors, treat as certainty units, scale variance using certainty units, and center using grand mean (equivalent to what SUDAAN does). If none of these options is acceptable for

variance estimation, the user can collapse strata to eliminate strata with only one PSU. The default is to report missing standard errors.

6.4. The R survey package provides multiple options for calculating survey degrees of freedom, depending on the user settings. When using degf(object), degrees of freedom are calculated as the number of PSUs minus the number of first stage sampling strata.

6.5. Sampling designs: R allows a variety of complex sampling designs including multistage, stratified, and clustered sampling with and without replacement. As with SUDAAN, any number of strata and sampling stages can be specified. When analyzing survey data, the sampling design information is packaged with the data into a 'survey design object', and this object is part of the input to analyses. For example, the YRBS design object is created as:

```
yrbsdes <- svydesign(id=~PSU, weight=~WEIGHT,
strata=~STRATUM,data=yrbs_data, nest=TRUE)
```

6.6. Sample program code:

```
install.packages("survey")
library (survey)
install.packages("haven")
library (haven)

yrbs_data <- read_sav("C:/YRBS2023/yrbs2023.sav")

yrbsdes <- svydesign(id=~PSU, weight=~WEIGHT, strata=~STRATUM,
data=yrbs_data, nest=TRUE)

options(digits=9)

seatbelt <- svyciprop(~I(QN8==1),yrbsdes, na.rm=TRUE, method = "xlogit")
seatbelt

eversex <- svyciprop(~I(QN56==1),yrbsdes, na.rm=TRUE, method = "xlogit")
eversex

heroin <- svyciprop(~I(QN52==1),yrbsdes, na.rm=TRUE, method = "xlogit")
heroin

seatbelt <- svymean(~I(QN8==1),yrbsdes, na.rm=TRUE, method = "xlogit")
seatbelt

eversex <- svymean(~I(QN56==1),yrbsdes, na.rm=TRUE, method = "xlogit")
eversex
```

```
heroin <- svymean(~I(QN52==1),yrbsdes, na.rm=TRUE, method = "xlogit")
heroin

seatbelt <- svytotal(~I(QN8==1),yrbsdes, na.rm=TRUE, method = "xlogit")
seatbelt

eversex <- svytotal(~I(QN56==1),yrbsdes, na.rm=TRUE, method = "xlogit")
eversex

heroin <- svytotal(~I(QN52==1),yrbsdes, na.rm=TRUE, method = "xlogit")
heroin

unwtd.count(~I(QN8==1), yrbsdes)
unwtd.count(~I(QN56==1), yrbsdes)
unwtd.count(~I(QN52==1), yrbsdes)

degf(yrbsdes)
```

Analysis commands in R do not typically produce all possible output immediately, instead they return an object that can be used to create further output if desired. For example, the output of svyciprop() does not include standard errors, but these can be extracted with the SE() function. Unweighted counts are not produced by default, but can be computed with the unwtd.count() function.

Cautionary note: When obtaining frequencies for a list of two or more variables, R uses list-wise exclusion and generates estimates using responses only from observations with no missing data for any of the variables on the list. This will be readily apparent from the unweighted sample sizes, which are all the same. To obtain proportions that use all non-missing data for each variable (i.e., pair-wise or table-by-table exclusion) the analysis must be run separately for each variable as shown above.

7. Comparison of Results

Table 3 shows results from seven different analyses of three variables from the 2023 national YRBS. For these comparisons, results on the table from some statistical software packages are shown to more decimal places than would be obtained by default. These analyses were run in the software versions that are listed in the column headings on Table 3; we would expect later versions of the same software would yield the same results. The first analysis – labeled naïve – is an unweighted analysis such as the one that would be obtained using SAS Proc Freq. This analysis is inappropriate for data from any YRBS and neither the point estimates nor the standard errors have been calculated correctly. The point estimates are not correct because the sampling weights were not used in the analysis; note that the values obtained do not match those from any of the analyses that used the sampling weights. In addition, the standard errors are about one-third to one-half as large as those obtained from an appropriate analysis since the variances were estimated using formulas that are appropriate for simple random sampling but not for complex samples. These formulas do not take stratification, clustering, or variability of sampling weights into account; this is important since both clustering and unequal sampling weights (unequal probabilities of selection) increase the variance. The resulting 95% confidence intervals are much narrower than those obtained from an appropriate analysis, giving the impression that the point estimates are more precise than they actually are. Because output from an analysis in SAS Proc Freq would not include standard errors and confidence intervals, these were obtained for this example using SAS Proc SurveyFreq with all three sample design statements omitted to simulate an analysis based on a simple random sample. Note that analyses in some general use statistical software packages (e.g., SPSS, Epi Info) may include standard errors or confidence intervals, even when the analyses performed were not appropriate for complex sample survey data. These standard errors or confidence intervals are invalid and should not be used.

The second analysis - a weighted analysis such as the one that would be obtained by including a WEIGHT statement in SAS Proc Freq – does result in valid point estimates, but the standard errors, while somewhat larger than those from the naïve analysis, are still much smaller than those obtained from an appropriate analysis of the data and the resulting 95% confidence intervals are still narrower than they should be. Note that this is the case even though national YRBS weights have been “standardized” to sum to the sample size rather than the population size. Because output from a weighted analysis in SAS Proc Freq would not normally include standard errors and confidence intervals, these were obtained for this example using SAS Proc SurveyFreq with only the WEIGHT design statement included to simulate a weighted analysis with no clustering or stratification. Note that weighted analyses in some general use statistical software packages (e.g., SPSS, Epi Info) may include standard errors or confidence intervals, even when the analyses performed were not appropriate for complex sample survey data. These standard errors or confidence intervals should not be used as they are invalid, even though the point estimates obtained from such weighted analyses are valid.

The next five analyses were performed using the complex sample survey procedures from the statistical software packages described in this document. Taylor Series Linearization

(TSL) was specified for variances estimation in SUDAAN and Stata, DESIGN = WR was specified in SUDAAN, and WR was selected for Stage 1: Estimation Method in SPSS. All analyses generate exactly the same point estimate for each variable since they are all doing a weighted analysis. In addition, the variances have been calculated taking stratification, clustering, and variability of sampling weights into account, resulting in larger standard errors and wider 95% confidence intervals than obtained in the two inappropriate analyses. There are some small differences in the standard errors and 95% confidence intervals obtained from analyses in these five statistical software packages, even though each was performing analyses that are appropriate for complex sample survey data. Since for these comparisons all variances were estimated using TSL, any observed differences in the standard errors are due to differences in the algorithms used by the different statistical software packages to perform the TSL, including differences in how single-PSU strata are handled for variance estimation. One reason for the differences in the confidence intervals, apart from differences in standard errors, is that some statistical software packages produce asymmetric (log transformed) confidence intervals for proportions while others produce symmetric confidence intervals. Another reason confidence intervals might differ between statistical software packages is the difference in how survey degrees of freedom are defined. A difference in survey degrees of freedom can be seen on Table 3 for the first variable analyzed. The observed differences in standard errors and confidence intervals are, however, quite small and inconsequential compared to the much larger differences in standard errors and confidence intervals between the appropriate and the inappropriate analyses. It would be appropriate to report the standard errors and confidence intervals shown on any of the last five lines of Table 3 for each variable.

8. Comparison of Statistical Software Packages

Each of the five selected statistical software packages described in this document can be used to analyze YRBS data appropriately; however, each of these statistical software packages has strengths and limitations. SUDAAN is not a general use statistical software package whereas SAS, Stata, SPSS, and Epi Info are. For analysts already using one of these four general use statistical software packages, a good option might be acquiring and using the sample survey modules or procedures available in the statistical software package they are already using. This includes the use of SAS-callable SUDAAN for SAS users. Epi Info is a no cost statistical software package available from the CDC to anyone. The other four are commercial statistical software packages which have varying annual licensing and update fees.

8.1. Analytic capabilities: SUDAAN, Stata, and R survey package have a wide range of analytic capabilities. All three have procedures for analyzing multiply imputed datasets, so that the variance due to multiple imputation can be included in the variance estimate and, in addition, Stata offers graphics and multi-level models with survey weights through its gllamm procedure; however, the R survey package offers the widest range of high quality graphics for complex survey data. SAS and SPSS are somewhat more limited in their analytic capabilities and Epi Info has very limited analytic capabilities. SUDAAN, SAS, Stata, and R estimate percentiles such as the median. SAS and SPSS do not offer a 2-sample t-test for comparison of domain means but these can be obtained using linear regression. Epi Info will calculate the difference in domain means and a confidence interval on the difference, but Epi Info does not offer a test of independence for crosstabulations and does not estimate population totals. SUDAAN, SPSS, and Stata produce asymmetric confidence intervals for proportions while SAS and Epi Info produce symmetric confidence intervals. R has multiple options for producing asymmetric confidence intervals for proportions. When obtaining frequencies for a list of two or more variables, Stata and R use list-wise exclusion and generates estimates using responses only from observations with no missing data for any of the variables on the list. To obtain proportions that use all non-missing data for each variable (i.e., pair-wise or table-by-table exclusion) the analysis must be run separately for each variable. The other four statistical software packages all use pair-wise (table-by-table) exclusion. Although all six statistical software packages allow estimation for domains, no subpopulation statement is available for analysis of subpopulations in SAS. Subpopulation analyses can be obtained by creating an indicator variable for subpopulation membership and using the indicator variable in a domain analysis. This also is true of subpopulation analyses in Epi Info. Subpopulation or domain estimates should never be obtained in SAS using the BY, IF, or WHERE statements since the variance estimates obtained in this manner will be invalid. All six statistical software packages can calculate design effect for proportions and all but Epi Info can calculate design effect for other estimated statistics.

8.2. Data requirements: SUDAAN and Stata require all variables used in an analysis to be numeric while SAS, SPSS, Epi Info, and R allow categorical variables to be either numeric or character. Additionally, R allows factor variables (variables that are represented as character strings in other packages would typically be represented as

factors in R). Only SUDAAN requires data to be sorted by the stratum and primary sampling unit (PSU) variables prior to analysis, although versions 10 and higher allow the user to put NOTSORTED on the procedure statement in lieu of sorting the data set. SUDAAN requires all variables to be coded >0, unless the variable is the dependent term in the logistic regression model.

8.3. Variance estimation: SPSS and Epi Info offer only Taylor Series Linearization (TSL) method of variance estimation while SAS, SUDAAN, Stata, and R offer balanced repeated replication and jackknife in addition to TSL. A finite population correction is not available for without replacement sampling designs in Epi Info, but is available in the other five statistical software packages. If an analysis includes data from one or more strata that contain only a single PSU, only SUDAAN will not proceed with the analysis. SUDAAN will run the analysis using the overall population mean to calculate the variance contribution for any singleton PSU if the MISSUNIT option appears on the PROC statement. SAS, SPSS, and Epi Info handle this situation by estimating the variance with no contribution from the single-PSU strata. Stata and R offer additional choices as to how the variance estimation will be handled in this situation, including those methods used by the other four statistical software packages.

8.4. Survey degrees of freedom: When analyzing variables for which data are missing for all respondents in one or more PSU or stratum, which most commonly occurs when performing analyses for a small subpopulation, the degrees of freedom will be calculated using the definition proposed by Korn and Graubard (1999) by all the described statistical software packages except SUDAAN. SUDAAN will overestimate the survey degrees of freedom and a remedy must be applied to obtain appropriate confidence intervals.

8.5. Sampling designs: SAS and Epi Info allow the sampling design to be specified at only the first stage of sampling. The other four statistical software packages allow the sampling design to be specified at additional stages of sampling. Thus SPSS is able to offer three sampling designs for variance estimation (WR, EqualWOR, and UnequalWOR) while SUDAAN, Stata, and R offer even more choices. SAS allows both with replacement (i.e., the first stage sampling fraction is negligible) and without replacement (i.e., use of a finite population correction by specifying a file containing stratum population totals for the first stage sampling units). Epi Info assumes a common sampling design for variance estimation which is the equivalent of specifying DESIGN = WR in SUDAAN.

8.6. Programming: Stata, SPSS, and Epi Info offer menu-driven (point-and-click) analysis but also can be run using syntax. Syntax must be written when using SAS, SUDAAN, or R.

Tables

Table 1 – Analytic Capabilities of Six Statistical Software Packages for Analysis of Complex Survey Data

Estimate or Analysis	SUDAAN 11.0 and higher	SAS 9.4 and higher	Stata 13 and higher	SPSS 21 and higher	Epi Info 7.2	R survey package 4.0 and higher
Means, proportions, Crosstabulations	X	X	X	X	X	X
Totals	X	X	X	X		X
Ratios	X	X	X	X		X
Median and other percentiles	X	X	X			X
Odds ratios, risk ratios	X	*	X	X	X	X
Odds differences					X	***
Test for difference in domain means	X	**	X	X	**	X
Chi-square tests	X	X	X	X		X
Linear regression	X	X	X	X		X
Logistic regression	X	X	X	X		X
Polychotomous logistic regression	X	X	X	X		****
Survival analysis	X	X	X	X		X
Poisson regression	X		X			X
Additional regression models			X			X
Design effect	X	X	X	X	X	X

* In SAS, odds ratios can be obtained using logistic regression.
** In SAS, a two-sample test for comparing means can be obtained using linear regression. Epi Info provides an estimated difference in domain means with standard error and confidence interval on the difference provided.
*** Odds differences using R survey package con be computed using svycontrast()
**** R survey package has capability to perform ordinal logistic regression but not unordered polytomous regression

Table 2 – Variance Estimation Methods of Five Statistical Software Packages for Analysis of Complex Survey Data

Variance Estimation Method	SUDAAN 11.0 and higher	SAS 9.4 and higher	Stata 13 and higher	SPSS 21 and higher	Epi Info 7.2	R survey package 4.0 and higher
Taylor Series Linearization (TSL)*	X	X	X	X	X	X
Balanced Repeated Replication (BRR)	X	X	X			X
Jackknife	X	X	X			X

* TSL is the default in statistical software packages with more than one variance estimation option available.

Table 3 – Results from Eight Analyses of 2023 National YRBS Data

YRBS variable Software/Procedure	Estimated %	Standard Error	95% Confidence Interval*	n†	Weighted n§	Survey degrees of freedom
Did not always wear a seat belt						
SAS v 9.4: Proc Freq¶	43.56	0.40	42.77, 44.35	15071	na	na
SAS v 9.4: Proc Freq weighted**	39.58	0.61	38.37, 40.79	15071	16917	na
SAS v 9.4:Proc SurveyFreq	39.58	1.51	36.56, 42.59	15071	16917	73
SUDAAN v 11.0: Proc Crosstab	39.58	1.72	36.21, 43.04	15071	16917	73††
SPSS v 21: CS Frequencies	39.58	1.72	36.21, 43.04	15071	16197	na
Stata v 18: Svy: Proportion	39.58	1.72	36.20, 43.05	15071	16917	63
Epi Info v 7.2: CSFrequencies	39.58	1.51	36.56, 42.59	15071	16917	na
R survey v. 4.2: svyciprop, svymean	39.58	0.02	36.21, 43.04	15071	16917	73
Ever had sexual intercourse						
SAS v 9.4: Proc Freq¶	33.62	0.37	32.89, 34.35	16121	na	na
SAS v 9.4: Proc Freq weighted**	31.58	0.58	30.44, 32.72	16121	17197	na
SAS v 9.4:Proc SurveyFreq	31.58	1.08	29.42, 33.74	16121	17197	73
SUDAAN v 11.0: Proc Crosstab	31.58	1.15	29.33, 33.91	16121	17197	73††
SPSS v 21: CS Frequencies	31.58	1.15	29.33, 33.91	16121	17197	na
Stata v 18: Svy: Proportion	31.58	1.15	29.33, 33.92	16121	17197	69
Epi Info v 7.2: CSFrequencies	31.58	1.08	29.42, 33.74	16121	17197	na
R survey v. 4.2: svyciprop, svymean	31.58	0.01	29.33, 33.91	16121	17197	73
Lifetime heroin use						
SAS v 9.4: Proc Freq¶	1.58	0.09	1.41, 1.76	19322	na	na
SAS v 9.4: Proc Freq weighted**	1.61	0.14	1.33, 1.88	19322	19365	na
SAS v 9.4: Proc SurveyFreq	1.61	0.44	0.73, 2.48	19322	19365	73
SUDAAN v 11.0: Proc Crosstab	1.61	0.44	0.93, 2.77	19322	19365	73††
SPSS v 21: CS Frequencies	1.61	0.44	0.93, 2.77	19322	19365	na
Stata v 18: Svy: Proportion	1.61	0.44	0.93, 2.77	19322	19365	73
Epi Info v 7.2: CSFrequencies	1.61	0.44	0.73, 2.49	19322	19365	na
R survey v. 4.2: svyciprop, svymean	1.61	<0.01	0.93, 2.77	19322	19365	73

Results from some statistical software packages are shown to more decimal places than would be obtained by default.

* Symmetric confidence intervals for proportions are obtained in SAS and Epi Info; asymmetric confidence intervals are obtained in SUDAAN, SPSS, and Stata.
† n = number of observations used in calculation (sample size).
§ Weighted n = weighted sample size (i.e., sum of weights for observations in the analysis rather than estimated total since national YRBS weights are standardized).
¶ Std errors and 95% confidence intervals obtained by simulating the analyses using SAS Proc SurveyFreq with no sample design statements. This method is naïve and produces incorrect estimates for prevalence, standard error, and confidence interval limits.
** Std errors and 95% confidence intervals obtained by simulating the analyses using SAS Proc SurveyFreq with only the Weight sample design statement. This method is naïve and produces incorrect estimates for standard error and confidence interval limits.
†† The SUDAAN degrees of freedom shown for Did not always wear a Seat Belt are not corrected.

Bibliography

Bell-Ellison, B and Kromrey, J (2007). Software Alternatives for Variance Estimation in the Analysis of Complex Sample Surveys: A Comparison of SAS Survey Procedures, SUDAAN, and AM, Proceedings of the Survey Research Methods Section, American Statistical Association, Alexandria, Virginia, 2659-2666.

Brogan, D (2007). Comparison of SAS and SUDAAN for BRFSS Descriptive Analyses. Workshop by Donna Brogan, Ph.D. 2007 Annual BRFSS Conference, Decatur, GA.

Brogan, D (2005). "Sampling Error Estimation for Survey Data", chapter in Household Sample Surveys in Developing and Transition Countries, editor-in-chief Graham Karlton, United Nations, New York, Section E, Chapter 21, pages 447-490.
website: http://unstats.un.org/unsd/HHsurveys/

Brogan, D (2005). "Software for Sample Survey Data: Misuse of Standard Packages", invited chapter in Encyclopedia of Biostatistics, 2nd Edition, editors-in-chief Peter Armitage and Theodore Colton, John Wiley & Sons, Ltd, Chichester, pages 5057-5064.

Brogan, D (1998). "Software for Sample Survey Data: Misuse of Standard Packages", invited chapter in Encyclopedia of Biostatistics, editors-in-chief Peter Armitage and Theodore Colton, John Wiley, New York, Volume 5, pages 4167-4174.

Carlson, BL (1998). "Software for Sample Survey Data", in Encyclopedia of Biostatistics, Volume 5 (of 6), edited by Peter Armitage and Theodore Colton. John Wiley & Sons, New York. 4160-4167.

Dean AG, Arner TG, Sunki GG, Friedman R, Lantinga M, Sangam S, Zubieta JC, Sullivan KM, Brendel KA, Gao Z, Fontaine N, Shu M, Fuller G, Smith DC, Nitschke DA, and Fagan RF (2011). Epi Info™, a database and statistics program for public health professionals. CDC: Atlanta, GA, USA. Available online at http://wwwn.cdc.gov/epiinfo/

Dorfman, A. and Valliant, R. (1993), "Quantile Variance Estimators in Complex Surveys," Proceedings of the Survey Research Methods Section, ASA, 866–871.

Francisco, C. A. and Fuller, W. A. (1991), "Quantile Estimation with a Complex Survey Design," Annals of Statistics, 19, 454–469.

Korn EL, Graubard BI. (1998) Confidence Intervals For Proportions With Small Expected Number of Positive Counts Estimated From Survey Data. *Survey Methodology* 23:193-201.

Korn, EL and Graubard, BI (1999). Analysis of Health Surveys. John Wiley & Sons, New York. p. 209-211.

Lumley T (2004) Analysis of complex survey samples. *Journal of Statistical Software* 9(1): 1-19.

Lumley T (2022) Package 'survey'. R package version 4.1-1. Available at https://cran.r-project.org/web/packages/survey/survey.pdf. Accessed 1/31/2023.

R Foundation for Statistical Computing (2009) *A language and environment for statistical computing.* Vienna, Austria. ISBN 3-900051-07-0.

Research Triangle Institute (2012). SUDAAN Language Manual, Volumes 1 and 2, Release 11. Research Triangle Park, NC: Research Triangle Institute.

Rust KF and Rao JNK (1996). Variance Estimation for Complex Surveys Using Replication Techniques. Statistical Methods in Medical Research, 5, 283-310.

SAS Institute, Inc. (2002-2004). SAS 9.1.3 Help and Documentation. Cary, NC: SAS Institute, Inc.

SAS Institute Inc. (2012). SAS OnlineDoc® 9.3. Cary, NC: SAS Institute Inc.

Särndal, C. E., Swensson, B., and Wretman, J. (1992), Model Assisted Survey Sampling, New York: Springer-Verlag.

Shah, BV (1998). "Linearization Methods of Variance Estimation", in Encyclopedia of Biostatistics, Volume 3 (of 6), edited by Peter Armitage and Theodore Colton. John Wiley & Sons, New York. 2276-2279.

Siller, AB and Tompkins, L (2005). "The Big Four: Analyzing Complex Sample Survey Data Using SAS®, SPSS®, STATA®, and SUDAAN®". Poster in 18th Annual Conference Proceedings, North East SAS Users Group (NESUG), September 11-14, 2005.

SPSS Inc. (2007). Introduction to SPSS Complex Samples™. Chicago, IL.

SPSS Inc. SPSS Complex Samples™ at http://www-03.ibm.com/software/products/en/spss-complex-samples.

StataCorp (2015). *Stata 14 Base Reference Manual*. College Station, TX: Stata Press

Survey Research Methods Section, American Statistical Association, Survey Software, http://www.hcp.med.harvard.edu/statistics/survey-soft/