In January 2009, CDC produced a report *Software for Analyis of YRBS data*, describing the use of SAS, SUDAAN, Stata, SPSS, and Epi Info for analyzing data from the Youth Risk Behaviors Survey.

This report provides the same information for R and the survey package. The text of section A is based on Section 3 of the CDC report, which describes Stata. Section B is based on Section 7 of the CDC report. It restricts itself to the features discussed in the CDC report and does not describe additional features of the R survey package. The tables extend those in the CDC report, providing two further analyses using R. The differences in results between R and SUDAAN are of the same size as the differences between the other packages, which the CDC report describes as "quite small and inconsequential".

R

R (http://www.r-project.org, R Foundation 2009) together with the R survey package (http://faculty.washington.edu/tlumley/survey, Lumley 2004, 2009) offers the ability to perform many statistical procedures on complex sample survey data, and graphics capabilities as well.

A.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, percentiles, 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. Domain estimates can be obtained using svyby(). Asymmetric confidence intervals for proportions are produced using svyciprop(), with options including "logit" for intervals that are symmetric on the log odds scale and "beta" for the method recommended by Korn & Graubard (1998). The survey package has facilities for analyzing multiply-imputed data, and other R packages such as 'mice' and 'mi' assist in creating these multiple imputations.

A.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.

A.3. Variance estimation: Variance estimation options available in R are Taylor Series Linearization (TSL) and replication methods including balanced repeated replication (BRR), jackknife, and bootstrap. 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.

A.4. Survey degrees of freedom: R 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 R. In addition, the preferred confidence interval estimation procedure for proportions (svyciprop(,method="beta")) does not require a 'degrees of freedom' estimate.

A.5. Sampling designs: Like SUDAAN, 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.

R, in contrast to Stata, SUDAAN, and SPSS, can have more than one data set in memory simultaneously. 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:

```
 yrbs <- svydesign(id=~psu, weight=~weight, strata=~stratum,
data=yrbs_data, nest=TRUE)
```

3.6 Sample code:

```
yrbs_data <- read.spss("yrbs07.sav",to.data.frame=TRUE)

yrbs <- svydesign(id=~psu, weight=~weight, strata=~stratum,
data=yrbs_data, nest=TRUE)

svyciprop(~I(QN8==1), yrbs, na.rm=TRUE)
svyciprop(~I(QN58==1), yrbs, na.rm=TRUE)
svyciprop(~I(QN52==1), yrbs, na.rm=TRUE)
```

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 is what usually should be intended, since it is the only approach that allows comparisons between variables in the output. However, it is not what CDC thinks is usually expected. 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.

Code for YRBS example, using SPSS input file:

```
yrbs_data <- read.spss("yrbs07.sav",to.data.frame=TRUE)

yrbs <- svydesign(id=~psu, weight=~weight, strata=~stratum,
data=yrbs_data, nest=TRUE)

## storing returned value in a variable
helmet <- svyciprop(~I(QN8==1), yrbs, na.rm=TRUE)
sex <- svyciprop(~I(QN58==1), yrbs, na.rm=TRUE)
heroin <- svyciprop(~I(QN52==1), yrbs, na.rm=TRUE)

helmet
sex
heroin
SE(helmet)
SE(sex)
SE(heroin)


## confidence intervals per Korn & Graubard (1998)
svyciprop(~I(QN8==1), yrbs, na.rm=TRUE,method="beta")
svyciprop(~I(QN58==1), yrbs, na.rm=TRUE,method="beta")
svyciprop(~I(QN52==1), yrbs, na.rm=TRUE,method="beta")

## unweighted counts
unwtd.count(~I(QN8==1), yrbs)
unwtd.count(~I(QN58==1), yrbs)
unwtd.count(~I(QN52==1), yrbs)
```

7. Comparison of Statistical Software Packages.

R, like SAS, STATA, SPSS, and Epi Info, is a general-purpose statistical package. Like Epi Info, it is available at no cost [unlike Epi Info, it is also open-source].

7.1 Analytic capabilities: R, like STATA and SUDAAN, has a wide range of analytic capabilities. It has procedures for analyzing multiply imputed datasets, so that the variance due to multiple imputation can be included in the variance estimate. R offers a widest range of graphics than any of the other packages.

Like SUDAAN and SAS, R will estimate percentiles such as the median. R offers a two-sample t-test for comparison of domain means, in addition to the ability to do this comparison with linear regression. R offers tests of independence in crosstabulations (and loglinear models for multiway tables).

Like SUDAAN, SPSS, and Epi Info, R has (multiple) options for producing asymmetric confidence intervals for proportions. Like Stata, 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. 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.

R can calculate all statistics for specific domains and across sets of subpopulations.

7.2. Data requirements: R allows numeric, character, and factor variables (variables that are represented as character strings in other packages would typically be represented as factors in R). There are no requirements on the sorting of data sets.

7.3. Variance estimation: R offers balanced repeated replication and jackknife in addition to Taylor Series Linearization, and provides for a finite population correction for sampling without replacement. R has the same choices as Stata for estimation when a stratum has only a single PSU.

7.4. 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, R will calculate degrees of freedom using the definition proposed by Korn and Graubard (1999), the same approach used by all the other packages except SUDAAN.

7.5. Sampling designs: R offers a full range of multistage sampling designs, and PPS sampling designs with and without replacement.

7.6. Programming: R does not provide a point-and-click GUI.

Table 1: Analytic capabilities

Means, proportions,crosstabulations	X
Totals	X
Ratios	X
Median and other percentiles	X
Odds ratios, risk ratios	X
Odds differences	*
Tests for difference in domain means	X
Chi-square tests	X
Linear regression	X
Logistic regression	X
Polytomous logistic regression	#
Survival analysis	X
Poisson regression	X
Additional regression models	X
Design effect	X

* svycontrast() can compute linear and nonlinear contrasts from survey output, including odds differences
# ordinal logistic regression, but not unordered polytomous regression.

Table 2: Variance estimation methods.

Taylor Series Linearization	X
Balanced Repeated Replication	X
Jackknife	X

Table 3: Results from analyses of YRBS data
(SUDAAN results taken from CDC report)

Analysis	Estimate	SE	CI	n	df
Never/rarely bike helmet					
SUDAAN 9.0 Proc crosstab	85.1033	1.3092	82.2592, 87.5604	8584	40
R svyciprop(,"logit")	85.1033	1.3092	82.3511, 87.4916	8584	39
R svyciprop(,"logit")	85.1033	1.3092	82.2372, 87.6676	8584	39
Ever had sexual intercourse					
SUDAAN 9.0 Proc crosstab	47.8423	1.3634	45.0960, 50.6016	13106	40
R svyciprop(,"logit")	47.8423	1.3634	45.1787, 50.5052	13106	39
R svyciprop(,"beta")	47.8423	1.3634	45.0551, 50.6395	13106	39
Lifetime heroin use					
SUDAAN 9.0 Proc crosstab	2.2635	0.2472	1.8143, 2.8207	13838	40
R svyciprop(,"logit")	2.2635	0.2472	1.8265, 2.8020	13838	40
R svyciprop(,"beta")	2.2635	0.2472	1.7907, 2.8206	13838	40

Additional References

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.

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

Lumley T (2009) "survey: analysis of complex survey samples". R package version 3.17.

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