

California Health Interview Survey



*california
health
interview
survey*

**Making
California's
Voices Heard
on Health**

California Health
Interview Survey
UCLA Center for Health
Policy Research
10960 Wilshire Blvd.
Suite 1550
Los Angeles,
California 90024

Tel (310) 794-0925
Toll Free (866) 275-2447
Fax (310) 794-2686

www.chis.ucla.edu

Weighting and Estimation of Variance in the CHIS Public Use Files

The California Health Interview Survey (CHIS) employs a two-stage geographically stratified random-digit-dial (RDD) sample design. This complex design requires proper weighting and variance (or its square root – standard error) calculation of the estimates. Most statistical software packages calculate variance assuming that the data are from a simple random sample; this underestimates the variance of estimates produced from the CHIS complex sample design. In order to accurately estimate variance without jeopardizing data confidentiality and respondent privacy, CHIS Public Use Files (PUFs) provide 80 replicate weights (`rakedw1`, ..., `rakedw80`) in addition to the final weight (`rakedw0`).

These weights fulfill different functions. The final weight (`rakedw0`) accounts for the sample selection probabilities and statistical adjustments for potential undercoverage and nonresponse biases. When this weight is applied, it ensures that estimates from the CHIS sample are an unbiased representation of the California population. The replicate weights (`rakedw1`, ..., `rakedw80`) are specially designed for valid variance estimation in the absence of the geographical sample design information (excluded from the CHIS PUFs). These 80 different weights provide variance estimates computed with 80 replications.

When using replicate weights in conjunction with the final weight, the estimates and their variance estimation are unbiased. When analyzing data from the CHIS PUFs, if the final weight is applied without the replicate weights unbiased estimates will be produced, but their variability will be underestimated due to the incorrect assumption that the sample is a simple random sample.

This document illustrates how the CHIS PUFs can be analyzed to produce valid variance estimates using SUDAAN and Stata V.9. These are two main software packages capable of incorporating replicate weights. The main difference in operating these software packages is that the sample design information is specified within each procedure for SUDAAN, whereas Stata requires sample design specification in a separate step preceding the analyses.

Sample code is provided for different types of analyses: for continuous variables, calculations of means and linear regression analysis are presented; and for categorical variables, calculations of frequencies and logistic regression analysis are presented. The estimates and their standard errors in all analyses are identical across the two software packages examined in this document.

For illustration purposes, Body Mass Index (`bmi_p`) is presented as a continuous dependent variable and current asthma status (`astcur`) as a categorical dependent variable. These variables are examined in relation to race (`racehpr`), sex (`srsex`) and age (`srage_p`). CHIS data users who wish to replicate the analysis presented here may copy the sample codes and generate the same results.

Example 1. Mean Calculation

In the sample code that follows, the distribution of BMI (`bmi_p`) is examined by race (`racehpr`) and by race and sex (`racehpr*srsex`).

SUDAAN:

```

PROC DESCRIPTIVE DATA = data FILETYPE = SAS DESIGN = JACKKNIFE;
WEIGHT rakedw0;
JACKWGTS rakedw1--rakedw80/adjjack=1;
VAR bmi_p;
TABLES racehpr racehpr*srsex;
SUBGROUP racehpr srsex;
LEVELS 7 2;
RUN;

```

Stata:

```

*Sample design specification step*a
use "DATASET LOCATION"
svyset [pw=rakedw0], jkrw(rakedw1-rakedw80, multiplier(1)) vce(jack) mse

*Analysis*
svy: mean bmi_p, over(racehpr)
svy: mean bmi_p, over(srsex racehpr)

```

^a = The sample design specification step should be included before conducting any analysis in Stata.

Example 2. Frequency Calculation

In the following sample code, the percentage of people who have asthma currently (astcur) is examined by race (racehpr) and by race and sex (racehpr*srsex).

SUDAAN:

```

PROC CROSSTAB DATA = data FILETYPE = SAS DESIGN = JACKKNIFE;
WEIGHT rakedw0;
JACKWGTS rakedw1--rakedw80/adjjack=1;
TABLES astcur*racehpr racehpr*astcur*srsex;
SUBGROUP astcur racehpr srsex;
LEVELS 2 7 2;
RUN;

```

Stata:

```
svy: tabulate astcur racehpr, col se ci
```

```
gen racesex=1 if racehpr ==1 & srsex ==1 a
replace racesex=2 if racehpr ==1 & srsex ==2
replace racesex=3 if racehpr ==2 & srsex ==1
replace racesex=4 if racehpr ==2 & srsex ==2
replace racesex=5 if racehpr ==3 & srsex ==1
replace racesex=6 if racehpr ==3 & srsex ==2
replace racesex=7 if racehpr ==4 & srsex ==1
replace racesex=8 if racehpr ==4 & srsex ==2
replace racesex=9 if racehpr ==5 & srsex ==1
replace racesex=10 if racehpr ==5 & srsex ==2
replace racesex=11 if racehpr ==6 & srsex ==1
replace racesex=12 if racehpr ==6 & srsex ==2
replace racesex=13 if racehpr ==7 & srsex ==1
replace racesex=14 if racehpr ==7 & srsex ==2
label variable racesex "Race x Sex"
label define racesexf 1 "latin male" 2 "latin fem" 3 "pi male" 4 "pi fem" 5 "na male" 6 "na
female" 7 "asian mal" 8 "asian fem" 9 "aa male" 10 "aa fem" 11 "white mal" 12 "white
fem" 13 "others mal" 14 "others fem"
label values racesex racesexf
```

```
svy: tabulate astcur racesex, col se ci
```

^a = Generating a new variable is recommended, as Stata creates overlapping categories when crossing variables.

Example 3. Linear regression

The following sample code examines Body Mass Index (bmi_p) in relation to race (racehpr), sex (srsex) and age (srage_p) while controlling for each other. Note that racehpr and srsex are categorical variables; and White (racehpr =6) and Male (srsex=1) are used as their reference categories.

SUDAAN:

```
PROC REGRESS DATA = data FILETYPE = SAS DESIGN = JACKKNIFE;
WEIGHT rakedw0;
JACKWGTS rakedw1--rakedw80/adjjack=1;
SUBGROUP racehpr srsex;
LEVELS 7 2;
REFLEVEL racehpr=6 srsex=1;
MODEL bmi_p = racehpr srsex srage_p;
RUN;
```

Stata:

```

recode racehpr (6=1) (1=2) (2=3) (3=4) (4=5) (5=6) (7=7), gen(race) a

xi: svy: regress bmi_p i.srsex i.race srage_p

```

^a = Recoding is done in order to choose “White” (racehpr=6) as the reference group

Example 4. Logistic regression

The following sample code examines current asthma status (astcur) among adults in California, controlling for race (racehpr), sex (srsex), and age (srage_p). As

SUDAAN and Stata require the dependent variables coded as 0 and 1 for logistic regression, a new dependent variable ast is created and assigned 1 where astcur=1 (“Current asthma”) and 0 where astcur=2 (“No current asthma”). The category, “No current asthma,” is used as the reference in the analysis.

SUDAAN:

```

DATA newdata;
SET data;
IF astcur=1 THEN ast=1;
ELSE IF astcur=2 THEN ast=0;
RUN;

PROC RLOGIST data = newdata FILETYPE = SAS DESIGN = JACKKNIFE;
WEIGHT rakedw0;
JACKWGTS rakedw1--rakedw80/adjjack=1;
SUBGROUP racehpr srsex;
LEVELS 7 2;
REFLEVEL racehpr = 6 srsex = 1;
MODEL ast = racehpr srsex srage_p;
RUN;

```

Stata:

```

recode astcur (2=0) (1=1) (-9=.), gen(ast)

xi: svy: logit ast srage_p i.race i.srsex a
xi: svy: logistic ast srage_p i.race i.srsex b

```

^a = This statement produces parameter estimates.

^b = This statement produces odds ratios.