Pooling CHIS Data Files from Multiple Years

The California Health Interview Survey (CHIS) was conducted as a biennial survey from 2001 through 2009. Beginning in 2011, CHIS data have been collected continuously across a two-year data collection cycle. Continuous data collection allows for the release of one-year data files and estimates for each calendar year. This document provides general guidelines for producing estimates and testing hypotheses with pooled/combined data from one-year public use adult data files (PUFs). SAS-callable SUDAAN® (SAS/SUDAAN) and Stata® codes that carry out specific tasks are provided below. You can download CHIS one-year public use data files here: http://healthpolicy.ucla.edu/chis/data/Pages/public-use-data.aspx

About the CHIS Design

CHIS is a representative sample of the California population using a two-stage, geographically-stratified dual-frame (cell phone and landline), random-digit-dial (RDD) sample. At the first stage, telephone numbers are drawn within 44 predefined geographic areas or "strata."[1] These telephone numbers are screened to determine if they are households and thus eligible for the survey. At the second stage, one adult is randomly selected from among the adults living the household. If there are any eligible adolescents or children in the household, one adolescent and/or child is selected for additional interviews. More detail about the sample design is available online in *Methodology Report 1: Sample Design* at: http://healthpolicy.ucla.edu/chis/design/Pages/methodology.aspx

Statistical Testing and Interval Estimation Overview

In order to assess the statistical stability of survey estimates, analytic results should always be presented with estimates of their variance (or its square root – the standard error). These often take the form of confidence intervals or margins of error. Another common measure of point estimate stability is the coefficient of variation (a point estimate divided by its standard error). In scientific publication, p-values and statistical tests may also be conducted. All of these methods are different ways to represent sampling variability (i.e., sampling variance) in statistical analyses, and they are all affected by the complex sample survey design in CHIS (see Analyze CHIS Data for more general information about special techniques for analyzing CHIS data: http://healthpolicy.ucla.edu/chis/analyze/Pages/default.aspx).

Estimate reliability/stability, confidence interval range/overlap, and p-values relative to alpha levels are all tools that researchers in various contexts use to make decisions and inferences about their results. For example, some organizations will not publish unreliable results as measured by the estimate's standard error, confidence interval, or coefficient of variation. Due to the complex sample design used in CHIS (geographically stratified sample with different probabilities of selection rather than a simple random sample of California), to accurately calculate variance estimates from CHIS data the survey's sample design must be taken into account.

CHIS PUFs contain weights that use a replication-based method for calculating valid standard errors for surveys with a complex sample design. For CHIS, the replicate weights `RAKEDW1` through `RAKEDW80`, as well as the final weight (`RAKEDW0`) are included in the PUF and must be used in analyses. The replicate weights correct variance estimates, and the final weight corrects point estimates. Thus, when all 80 replicate weights are applied, variances will be estimated correctly.

[1] In earlier years the number of strata has ranged from 41 in 2001 to the current number, 44 in 2013-2014.

Replicate Variance Estimation Implementation in CHIS

This section covers how to analyze pooled multi-year CHIS data and produce appropriate variance and point estimates. SAS®, SUDAAN® and Stata® (Version 9 and higher) are commonly- used software packages with features to handle survey data with replicate weights. We will use SAS, SAS-callable SUDAAN® (Research Triangle Institute, 2004) and Stata® (StataCorp., 2005) in the following sections.

The sample code provided below assumes that users will combine either a CHIS one-year data set with another one-year data set, a CHIS two-year data set with another CHIS two-year data set, or a one-year data set with a two-year data set. CHIS does not recommend pooling continuous data (CHIS 2011 and beyond) with CHIS data collected prior to 2010 due to methodological changes that affect the comparability of data collected before and after the 2010 U.S. Census. For more information on methodological changes in CHIS over cycles, please refer to the CHIS Methodology website: http://healthpolicy.ucla.edu/chis/design/Pages/methodology.aspx.

A sketch of constructing statistical weights for pooling two one-year data files is provided in Table 1. Code that can be used to create the pooled datasets and the weight variables needed for the correct estimation of standard errors can be found below. To create a file for this example, we concatenated the 2013 and 2014 Public Use Files (i.e., append the 2014 file to the 2013 file to create a single data file). The number of respondents in the combined data file is the sum of the respondents in the two individual data files. There are two main tasks to be carried out to create the combined data file:

1) Variables that will be used in the analyses should have the same name and categories in both data files. For example, make sure that education in both files has four categories that mean the same thing. This is something a data user will need to confirm independently.

2) Create a new set of weights that takes the two-year files into account.

To accomplish step two, refer to Table 1. Your new two-year replicate weight variables will include 161 weights in the combined data file: one final weight and 160 replicate weights (80 for 2013 and 80 for 2014). Note that you only need one final weight, but you need twice as many replicate weights as were included in the PUFs. The final weight in the combined file is created using the final weight (`RAKEDW0`) from each of the two data files. The creation of the 160 replicate weights is more complex.

1) To generate the final weight for the CHIS 2013 cases in your pooled file, assign your new final weight variable the value of `RAKEDW0` (final weight) from the CHIS 2013 data and divide it by 2.

2) To generate the 160 replicate weights for CHIS 2013 cases,
   a. Assign the first 80 replicate weights to be the same values as the original CHIS 2013 replicate weights `RAKEDW1` through `RAKEDW80` and divide each by 2.
   b. Assign each of the new replicate weights 81-160 **for CHIS 2013 cases**, to the value of the CHIS 2013 final weight (`RAKEDW0`), and divide each by 2. All 80 weight variables will have the same value.

3) To generate the final weight for the CHIS 2014 cases, assign the new final weight variable the value of the CHIS 2014 `RAKEDW0` (final weight) and divide it by 2. Your new final weight variable will now be assigned the CHIS 2013's `RAKEDW0` value for the CHIS 2013 cases and the CHIS 2014's `RAKEDW0` value for the CHIS 2014 cases.

4) To generate the 160 new replicate weights for the CHIS 2014 cases, reverse the assignment process from Step 2):
    a. Assign the first 80 replicate weights, the value of RAKEDW0 and divide each by two.
    b. Assign replicate weights 81-160 to be equal to the values of the original CHIS 2014 replicate weight variables RAKEDW1–RAKEDW80 and divide each by 2.

An Explanation of the Weights to Motivate the Approach

Within each one-year data set, the final weight, RAKEDW0, reflects the number of Californians each respondent represents in the data – for example, a case with a weight of 2355 means that the respondent (and their answers) represents 2355 Californians. Thus, the sum of RAKEDW0 across all age groups is an estimate of the total California population based on the control totals used for this survey. You can check this number against California Department of Finance or Census Bureau estimates for the same time period, but you should not expect it to match exactly.

For the purposes of pooling, to ensure that the population estimates and standard errors reflect the **average** California population over the pooled two-year period, the final weight and each replicate weight must be divided by 2 (as discussed in the steps above and displayed in Table 1 below).

Table 1. Construction of Statistical Weights for the Combined Data File

Year	Final weight	Replicate weight 1-80	Replicate weight 81-160
1 (CHIS2013)	**CHIS 2013:** final weight (2013's RAKEDW0/2)	**CHIS 2013:** replicate weights (2013's RAKEDW1/2, …, RAKEDW80/2)	**CHIS 2013:** final weight (2013's RAKEDW0/2 repeated 80 times)
2 (CHIS 2014)	**CHIS 2014:** final weight (2014's RAKEDW0/2)	**CHIS 2014:** final weight (2014's RAKEDW0/2 repeated 80 times)	**CHIS 2014:** replicate weights (2014's RAKEDW1/2, …, RAKEDW80/2)

Sample Code to Pool Multiple Cycles of CHIS Data

Visit the Analyze CHIS Data page to access SAS-callable SUDAAN® (SAS/SUDAAN) and Stata® sample codes to concatenate data files from multiple years and generate a new set of weights that take multi-year pooling into account: http://healthpolicy.ucla.edu/chis/analyze/Pages/sample-code-pooling.aspx

Replicate Weight Adjustment When Pooling One-Year Data with Two-Year CHIS 2011-12:

When pooling the one-year CHIS 2013 data with the two-year CHIS 2011-12 dataset, the final weight must be adjusted to account for the fact that the population estimates for the CHIS 2011-12 are weighted to reflect the total California population for 2012, but contain data collected over the full two-year period. Dividing the final weight by two in this instance would result in giving estimates from the 2013 file twice the weight of those collected in either 2011 or 2012. Instead, the code assigning fnwgt0 = rakedw0/2 should be changed to:

```
fnwgt0 = rakedw0*2/3 in the CHIS 2011-2012 data; and
fnwgt0 = rakedw0*1/3 in the CHIS 2013 data
```

In addition, each of the replicate weights in the CHIS 2011-2012 data should be multiplied by 2/3 rather than divided by 2; while each of the replicate weights in the CHIS 2013 data should be multiplied by 1/3 rather than divided by 2 in the CHIS 2013 data. Making this adjustment will lead to population estimates that give equal weight to each year of data.

For more information about analyzing the combined data file for estimating means, frequencies and their changes over time and calculating changes over time using linear regression and logistic regression, please see: http://healthpolicy.ucla.edu/chis/faq/Documents/paper_trends_averages.pdf

References

Research Triangle Institute. (2004). *SUDAAN Example Manual: Release 9.0*. Research Triangle Park, NC: Research Triangle Institute.

StataCorp. (2005). *Stata Statistical Software: Release 9.0.* College Station, TX: StataCorp.