https://healthpolicy.ucla.edu/chis/analyze/Pages/sample-code-pooling.aspx

Pooling CHIS Data

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. The following sections provide general guidelines for producing estimates and testing hypotheses with pooled/combined data from one-year public use files (PUFs). To download these guidelines, please refer to Pooling CHIS Data.

You can download CHIS one-year public use data files here:
http://healthpolicy.ucla.edu/chis/data/Pages/public-use-data.aspx

Sample Code to Pool Multiple Cycles of CHIS Data can be found here:
https://healthpolicy.ucla.edu/chis/analyze/Pages/sample-code-pooling.aspx

Additionally, a SAS macro for users interested in analyzing multi-year CHIS data to create adjusted replicate weight variables can be downloaded here:
https://healthpolicy.ucla.edu/chis/analyze/Pages/sample-code-pooling.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.

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.

Replicate Variance Estimation Implementation in CHIS

This section covers how to analyze pooled multi-year CHIS data and produce appropriate variance and point estimates.

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.

We will like to note that variables that will be used in the pooled-year analyses should have the same name and categories in all pooled CHIS year data files. For example, make sure that education in CHIS 2018 and CHIS 2019 files has four categories that mean the same thing while pooled CHIS 2018 and 2019 data. This is something a data user will need to confirm independently.

The sketch of constructing replicate weights for pooling two one-year data files are provided in Table 1. Codes 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 2018 and 2019 Public Use Files (i.e., append the 2019 file to the 2018 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 with new set of weights that takes the two pooled one-year files into account, 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 2018 and 80 for 2019). Note that you only need one final weight, but you need twice as many replicate weights as were included in the PUFs.

Step 1. To generate the final weight for the CHIS 2018 cases in your pooled file, assign your new final weight variable the value of RAKEDW0 (final weight) from the CHIS 2018 data and divide it by the number of one-year CHIS data files, which is 2 here.

Step 2. To generate the 160 replicate weights for CHIS 2018 cases, assign the first 80 replicate weights to be the same values as the original CHIS 2018 replicate weights RAKEDW1 through RAKEDW80 and divide each by 2. Assign each of the new replicate weights 81-160 for CHIS 2018 cases, to the value of the CHIS 2018 final weight (RAKEDW0), and divide each by 2. All 80 weight variables will have the same value. Similar process has to be repeated to generate the final weight for the CHIS 2019 cases, assign the new final weight variable the value of the CHIS 2019 RAKEDW0 (final weight) and divide it by 2. By that, the new final weight variable will now be assigned the CHIS 2018's RAKEDW0 value for the CHIS 2018 cases and the CHIS 2019's RAKEDW0 value for the CHIS 2019 cases.

An Explanation of 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).

A general rule of thumb is:

- The final number of replicate weights created after pooling equals the number of datafiles used (regardless of whether it is a one-year datafile or two-year file) times 80.

  For example, pooling CHIS 2018 and CHIS 2019 will result in $2 \times 80 = 160$ replicate weights.

  Pooling CHIS 2017-2018 two-year data with CHIS 2019 one-year will also create $2 \times 80 = 160$ replicate weights.

- The proportion each datafile will take in final base and replicate weights depends on the number of year(s) it represents, i.e., a one-data file takes one portion, and a two-year datafile take two portions.

  For example, pooling CHIS 2018 and CHIS 2019 (both one-year files) will create a dataset representing two years of data. For each year, the final base and replicate weights will be 1/2 of the original base and replicate weight.

  Pooling CHIS 2017-2018 and CHIS 2019: the final data should represent three years of data, and CHIS 2017-2018 should take two portions (2/3) and CHIS 2019 takes one portion (1/3).

Table 1. Construction of Statistical Weights for the Combined Data File (e.g., CHIS 2018 and 2019):

Year	**Final Weight**	**Replicate Weight 1-80**	**Replicate Weight 81-160**
CHIS 2018	RAKEDW0/2	RAKEDW1/2,…, RAKEDW80/2	RAKEDW0/2
CHIS 2019	RAKEDW0/2	RAKEDW0/2	RAKEDW1/2,…, RAKEDW80/2

Replicate Weight Adjustment When Pooling One-Year Data with Two-Year CHIS Data
When pooling the one-year data with the two-year dataset, the final weight must be adjusted to account for the fact that the population estimates for the two-year data are weighted to reflect the total California population for the middle of the two-year, 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 one-year file twice the weight of those collected over the two-year cycle.

Using 2017-2018 and 2019 as an example:

Instead, the code assigning fnwgt0 = rakedw0/2 should be changed to:
fnwgt0 = rakedw0*2/3 in the CHIS 2017-2018 data; and
fnwgt0 = rakedw0*1/3 in the CHIS 2019 data

In addition, each of the replicate weights in the CHIS 2017-2018 data should be multiplied by 2/3 rather than divided by 2; while each of the replicate weights in the CHIS 2019 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.