Examining Trends and Averages Using Combined Cross-Sectional Survey Data from Multiple Years

Sunghee Lee, UCLA Center for Health Policy Research
William W. Davis, National Cancer Institute
Hoang Anh Nguyen, UCLA Center for Health Policy Research
Timothy S. McNeel, Information Management Services, Inc.
J. Michael Brick, Westat
Ismael Flores-Cervantes, Westat

This document provides general guidelines for producing estimates and hypothesis tests using data from two cross-sectional surveys. These methods are illustrated using data from the California Health Interview Survey (CHIS) Public Use Files (PUF). Sample programming code in SAS-callable SUDAAN (SAS/SUDAAN) and Stata to implement the specific tasks described in this document are provided in the Appendix.

Introduction

CHIS is a collaborative project of the UCLA Center for Health Policy Research, the California Department of Health Services, and the Public Health Institute. The survey has been conducted biennially since 2001 using a two-stage, geographically stratified random-digit-dial (RDD) sample design to produce a representative sample of the state. At the first stage, telephone numbers are drawn within predefined geographic areas or "strata."[1] 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 among all adults living in a household. If there are any eligible adolescents or children in the household, one adolescent and/or child is selected for the additional interviews. Westat, a research corporation, has performed data collection services for each CHIS cycle. This paper uses the CHIS 2001 and 2003 adult data files to illustrate how to combine and use data from multiple CHIS cycles.

To assess the statistical stability of survey estimates, analytic results are presented with their variance estimates (or square roots – the standard errors). Valid variance estimation must take into account the complex sample design of the survey, otherwise the variance will be underestimated. Replication is a technique for calculating valid standard errors for surveys using a complex sample design without revealing the sample design information, which may jeopardize survey respondents' privacy and data confidentiality. For CHIS, replication weights (RAKEDW1, …, RAKEDW80) as well as a final weight (RAKEDW0) are included in the Public Use Files (PUFs). While applying the final weight alone produces an unbiased estimate, all 80 replication weights should be applied to estimate the variances correctly. For more information on the CHIS sample design, weighting, and estimation, see *"Methodology Brief - Weighting and Estimation of Variance in the CHIS Public Use Files"* at http://www.chis.ucla.edu/methods.html.

WesVar, 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-callable SUDAAN (Research Triangle Institute, 2004) and Stata (StataCorp., 2005) in this document.

An example of estimates from CHIS 2001 and CHIS 2003

Table 1 summarizes the CHIS 2001 and 2003 weighted estimates of the percentage of women ages 40 and older who had a mammogram in the last two years for the entire

---

[1] The number of strata increased to 44 in CHIS 2005.

sample and for subgroups of age, race/ethnicity, urban/rural status, and education. These results were separately obtained using the 2001 and 2003 data files. The standard errors account for the complex sample design by using the replicate weights. For both years, the largest standard error is found for the Native Hawaiian & Pacific Islander group, reflecting its small sample sizes.

Table 1. Mammogram Usage for Women 40+ from CHIS 2001 and 2003

Characteristics	2001		2003	
	%	SE	%	SE
All	75.38	0.39	76.08	0.45
Age				
40-49	66.21	0.66	67.60	0.93
50-64	82.39	0.65	82.30	0.62
≥ 65	78.38	0.74	78.75	0.92
Race/ethnicity				
Latino	69.45	1.16	69.96	1.39
Native Hawaiian & Pacific Islander	59.15	8.69	82.58	6.48
American Indian & Alaska Native	69.85	3.94	69.19	4.83
Asian	67.98	1.47	73.99	1.61
African American	77.53	1.60	77.16	1.73
White	78.41	0.46	78.22	0.55
Other/Multiple race	71.37	3.04	76.40	2.40
Metropolitan status				
Urban	75.41	0.41	76.24	0.47
Rural	74.60	0.84	72.30	1.36
Education				
<High school	68.31	1.27	69.52	1.36
High school	74.92	0.78	75.84	0.91
Some college +	78.22	0.46	78.45	0.55

Research based on a series of cross-sectional surveys often emphasize the results of the new survey but also include trend results. This document focuses on four general goals and provides examples for each when making inferences from multiple cross-sectional survey:

Goal 1.   Estimating a change in a characteristic such as a mean or a percentage and testing the significance of the change:
   Example 1:   Has the percentage of women ages 40 and older who had a mammogram in the last two years changed? What is the estimate of the change?
   Example 2:   Has mammogram usage changed for the Native Hawaiian & Pacific Islander subgroup? What is the estimate of the change?
Goal 2.   Estimating a change in a characteristic controlling for covariates:
   Example 1:   Has mammogram usage changed for the Native Hawaiian & Pacific Islander subgroup when controlling for age, education, and metropolitan status?
   Example 2:   Is there a difference by race/ethnicity in mammogram usage change when controlling for the other factors?
Goal 3.   Estimating the average using data from multiple survey years assuming that the mean has not changed between those years:
   Example: What is the average mammogram usage estimate for 2001 and

2003?

Goal 4. Estimating the population total using data from multiple survey years:

        Example: What is the total number of women with mammogram usage for 2001 and 2003?

Goal 1. Estimating changes - without controlling for other factors

It is simple to produce an estimate of change in characteristics between 2001 and 2003 and its corresponding variance estimate, because CHIS samples are drawn independently. We call CHIS 2001 year 1 and CHIS 2003 year 2 and consider estimating a characteristic θ (mean or percentage) in year s. We label the true (unknown) value in year s as θ_s, the estimated value as $\hat{\theta}_s$, and the estimated variance (the square of the standard error) as $v(\hat{\theta}_s)$. These estimates can be computed from data files from each year without pooling data files from the two years. The true change is

$$\Delta = \theta_2 - \theta_1,\tag{1}$$

where (1) is estimated by $\hat{\Delta} = \hat{\theta}_2 - \hat{\theta}_1$. Since CHIS 2001 and 2003 use independent samples, the estimated variance of the change, $v(\hat{\Delta})$, is

$$v(\hat{\Delta}) = v(\hat{\theta}_2) + v(\hat{\theta}_1).\tag{2}$$

Table 2 provides a summary of this information.

Table 2. Summary of Estimating Changes Using Two Surveys

Year	True value	Estimated value	Variance of estimate
1	θ_1	$\hat{\theta}_1$	$v(\hat{\theta}_1)$
2	θ_2	$\hat{\theta}_2$	$v(\hat{\theta}_2)$
Change	$\Delta = \theta_2 - \theta_1$	$\hat{\Delta} = \hat{\theta}_2 - \hat{\theta}_1$	$v(\hat{\Delta}) = v(\hat{\theta}_2) + v(\hat{\theta}_1)$

A hypothesis test for no change can be obtained by dividing $\hat{\Delta}$ by its standard error, $s(\hat{\Delta}) = \sqrt{v(\hat{\Delta})}$, and calculating a *p*-value using the Student-*t* distribution. Alternatively, the hypothesis can be tested at the 5% level by computing a 95% confidence interval (CI) and rejecting the hypothesis if the CI does not include 0. The confidence interval is computed as follows:

$$CI(\hat{\Delta}) = \hat{\Delta} \pm t \times s(\hat{\Delta})\tag{3}$$

where *t* is a percentage point of the Student-*t* distribution with degrees of freedom equal to the total number of replicate weights for the two surveys combined. Since there are 80 replicate weights, *t* for computing 95% confidence intervals in this case will be approximately 1.96. Using the total number of replicate weights as default degrees of freedom for the *t* value may not be appropriate and should be evaluated in specific applications (Westat, 2005). However, computing exact degrees of freedom for a complex survey like CHIS is very difficult. As long as the number of replicates is large and the distribution of the respondents is not highly concentrated in a few geographic areas, making the normal distribution assumption is reasonable.

Table 3 shows the results of estimating change from 2001 to 2003 in the percentage of women age 40+ who had a mammogram in the last two years. We can test the hypothesis of no change between 2001 and 2003 at the 5% significance level using the confidence interval approach. Not surprisingly, the table shows that most of the confidence intervals include 0, as one would not expect dramatic changes in a two-year period. The only confidence intervals that do not include 0 are for the Native Hawaiian & Pacific Islander subgroup and the Asian subgroup, both indicating a significant increase in mammogram usage. Note that these results are unadjusted for the impact of the other factors. The adjusted results will be discussed when we consider Goal 2.

Table 3. Mammogram Usage for Women 40+: Changes from CHIS 2001 to 2003

Characteristics	Estimated change (%)	95% CI Lower bound	95% CI Upper bound
All	0.70	-0.48	1.89
Age			
40-49	1.39	-0.86	3.64
50-64	-0.08	-1.85	1.69
≥ 65	0.38	-1.94	2.69
Race/ethnicity			
Latino	0.50	-3.08	4.08
Native Hawaiian & Pacific Islander	23.43	2.04	44.83
American Indian & Alaska Native	-0.66	-12.97	11.65
Asian	6.01	1.70	10.32
African American	-0.36	-5.02	4.30
White	-0.20	-1.61	1.22
Other/Multiple race	5.02	-2.62	12.67
Metropolitan status			
Urban	0.83	-0.41	2.06
Rural	-2.31	-5.47	0.86
Education			
<High school	1.21	-2.46	4.88
High school	0.92	-1.45	3.29
Some college +	0.23	-1.19	1.65

The method described above may be preferred for estimating change due to its simplicity. (It does not require combining the two data files.) Alternatively, the changes between years can be estimated using a combined data file. The combined data approach is especially useful when the goal of analysis is more than just calculating the unadjusted change over time. We outline how to pool the two data files into a single file and show the alternative method for estimating change using the combined data. The method using the combined data gives the same results as the method of using separate data files for calculating change.

In order to create the combined data file, concatenate the 2001 and 2003 files so that the number of respondents in the combined data file is the sum of the respondents from the two individual data files. Two main tasks are required to combine data files. First, variables used in the analyses should have the same name and values or categories in both data files. Section A of the Appendix describes how variables are redefined for the

tasks in this document. Second, create a set of new statistical weights as shown in Table 4. There will be 161 weights in the combined data file: one final weight and 160 replicate weights. We label them `FNWGT0` and `FNWGT1` to `FNWGT160`. The final weight (`FNWGT0`) in the combined file is created by using the final weight (`RAKEDW0`) from the respective surveys. For the first 80 replicate weights (`FNWGT1, …, FNWGT80`), we use replicate weights (`RAKEDW1, …, RAKEDW80`) for CHIS 2001 sample persons and final weights (`RAKEDW0`) for the CHIS 2003 sample. CHIS 2003 samples will have the same value for the first 80 replicate weights. The construction pattern is reversed for replicate weights 81-160 (`FNWGT81, …, FNWGT160`), as the value of same weight (`RAKEDW0`) is given to all 81-160 replicates for the CHIS 2001 sample while the actual replicate weights (`RAKEDW1, …, RAKEDW80`) are used for the 2003 sample. In Appendix Section B.1, SAS/SUDAAN and Stata code to combine the data files and create the weights listed in Table 4 are given. Following the same logic, combining data files from three years will require 241 weights, where one is the final weight and the remaining 240 are the new replicate weights as given in Appendix Section B.2. The general design statements for using combined data are provided in Appendix Section C.

Table 4. Construction of Statistical Weights for the Combined Data File

Combined Data	Combined Data: Final weight (FNWGT0)	Combined Data: Final replicate weight 1-80 (FNWGT1, …, FNWGT80)	Combined Data: Final replicate weight 81-160 (FNWGT81, …, FNWGT160)
Year 1 Sample (CHIS 2001)	CHIS 2001: Final weight (RAKEDW0)	CHIS 2001: Replicate weights (RAKEDW1, …, RAKEDW80)	CHIS 2001: Final weight (RAKEDW0, …, RAKEDW0)
Year 2 Sample (CHIS 2003)	CHIS 2003: Final weight (RAKEDW0)	CHIS 2003: Final weight (RAKEDW0, …, RAKEDW0)	CHIS 2003: Replicate weights (RAKEDW1, …, RAKEDW80)

Now we describe how to obtain estimates equivalent to those given in Table 3 using the combined data and the new set of weights defined above. The parameter $\Delta = \theta_2 - \theta_1$ of equation (1) is a contrast term since it can be expressed as a linear combination of the basic parameters θ_1 and θ_2, where the sum of their coefficients is 0. The coefficients are 1 for individuals sampled in year 2 and -1 for those sampled in year 1. Since Δ is the difference between two parameters, it can be estimated in SAS/SUDAAN using the `DIFFVAR` (or `CONTRAST`) option within `PROC DESCRIPT` or in Stata using the `LINCOM` command (see example codes in Appendix Section D).

Goal 2. Estimating changes – controlling for other factors

We provide a single method for estimating change while controlling for other factors using the combined data. Suppose, for example, that we are interested in determining whether the changes in mammogram usage for the Native Hawaiian & Pacific Islander group and the Asian group from Table 3 are statistically significant when we control for age, metropolitan status, and education.

There is less agreement on the best approach as analytic goals become more complicated; however, we suggest the following:
- Use the combined file with the new weights from Table 4; and
- Use a regression approach (multiple regression for numerical outcomes and logistic regression for binary outcomes) that includes the survey year as one of the covariates, assuming that the effect of the other covariates is the same across years.

To test if the change remains significant for a particular group, we can carry out a regression analysis restricting the data set to that group. If we use the first year as a reference category of the survey year covariate, the estimated coefficient for the second year can be used to determine the statistical significance of the change. The result of the significance test can be determined in the following ways:
- Examine whether a 95% confidence interval for the beta coefficient for the survey year contains 0; or
- Examine whether the *p*-value of the Wald F test for the survey year is less than 0.05.

Table 5 shows the effect of year in mammogram usage for Native Hawaiian & Pacific Islander and Asians when controlling for age, metropolitan status, and education (coefficients of the control variables not shown). This result comes from two separate logistic regression analyses using only the Native Hawaiian & Pacific Islander subgroup and only the Asian subgroup. The confidence intervals for the beta coefficients do not contain 0, and the *p*-values are less than 0.05. Therefore, it can be concluded that there were significant increases in mammogram usage between CHIS 2001 and 2003 for these two groups, even after controlling for age, urban/rural status, and education. This analysis can be carried out with PROC RLOGIST in SAS/SUDAAN or SVY:LOGIT in Stata (see Appendix Section E).

Table 5. Changes in Mammogram Usage for Women 40+ between 2001 and 2003 for Native Hawaiians & Pacific Islanders and Asians Controlling for Age, Metropolitan Status, and Education

Effect of year	β coefficient (95% CI)	*p*-value
Native Hawaiian & Pacific Islander		
2001	0.00	---
2003	1.40 (0.09, 2.71)	0.036
Asian		
2001	0.00	---
2003	0.25 (0.03, 0.46)	0.027

Recall that Table 3 also suggested that the change in mammogram usage for these two subgroups were different from that of whites: these two groups showed a significant increase in mammogram usage over time, while whites showed no change. However, this conclusion was based on an unadjusted analysis. To test whether the conclusion remains when taking into account other covariates, we use a logistic regression model with an interaction between race/ethnicity and the survey year. This interaction term is equivalent to the difference in change among racial/ethnic groups. A similar model was used by Korn and Graubard (1999, Sec. 8.4) for testing whether education levels had a differential effect on change in mammogram usage rates when adjusting for other factors using the National Health Interview Survey data.

To examine the differences in mammogram usage change over time among various racial/ethnic groups, our analysis uses whites as the reference group. Appendix Section F shows the corresponding programming code for SAS/SUDAAN and Stata. The beta coefficients in Table 6 represent the difference of the change on the logit scale between the Native Hawaiian & Pacific Islander subgroup and the reference group (i.e., whites) and between the Asian subgroup and the whites (coefficients of other variables and categories not shown). The results indicate that the change in mammogram usage for the two groups is different from the change for whites when controlling for other factors. (The odds ratio of the interaction effect is difficult to interpret and thus is not reported.) One difference between Tables 5 and 6 is that the estimates from Table 5 are based on data only with a specific race/ethnicity subgroup, while the results of Table 6 consider the entire sample.

Table 6. Differences in Changes for Native Hawaiians & Pacific Islanders and Asians Compared to Changes for Whites in Mammogram Usage for Women 40+ Between 2001 and 2003, Controlling for Age, Metropolitan Status, and Education

Effect of interaction between race and year	β coefficient (95% CI)	*p*-value
Native Hawaiian & Pacific Islander		
2001	0.00	---
2003	1.22 (0.06, 2.38)	0.039
Asian		
2001	0.00	---
2003	0.26 (0.02, 0.50)	0.032

Goal 3. Estimating average using two years of data

With two distinct surveys, we may report separate values for two surveys or one value summarizing the entire time period (e.g., average). If the distinct estimates from the two years are quite different, then reporting their average may not be a good idea, since the average may represent neither of the surveys. The significance test used for Goal 1 could serve to determine whether to report two distinct values or a single value, but practical consideration of the difference might be a better criterion.

The average of two survey years may be estimated in two ways: 1) using two separate data files and 2) using the combined data file. In the first approach, we use the mean value $\theta_m = 0.5 \times (\theta_2 + \theta_1)$ as the parameter of interest. Table 7 shows how we would estimate the mean and its variance. In general terms, the following estimators can be used for θ_m and $v(\theta_m)$ when combining data files from multiple cycles: $\hat{\theta}_m = \dfrac{\sum_{s=1}^{S} \hat{\theta}_s}{S}$ and

$v(\hat{\theta}_m) = \dfrac{\sum_{s=1}^{S} v(\hat{\theta}_s)}{S^2}$, where $\hat{\theta}_s$ is the estimate from year s with $s = 1,...,S$ and S is the number of years whose data are combined. The means calculated with this technique are reported in Table 8 under the heading "Mean," where the values for the individual years are shown in Table 1. Note that this is the method that *Ask*CHIS, the online query system that provides CHIS estimates, uses to calculate estimates when combining multiple years. Standard errors from *Ask*CHIS are likely to differ from the ones reported in this document, as the analyses in *Ask*CHIS use SAS with sample design information rather than replicate weights. However, the differences are small.

Table 7. Summary of Estimating Means Using Two Surveys

Year	True value	Estimated value	Variance of estimate
1	θ_1	$\hat{\theta}_1$	$v(\hat{\theta}_1)$
2	θ_2	$\hat{\theta}_2$	$v(\hat{\theta}_2)$
Mean	$\theta_m = 0.5 \times (\theta_1 + \theta_2)$	$\hat{\theta}_m = 0.5 \times (\hat{\theta}_1 + \hat{\theta}_2)$	$v(\hat{\theta}_m) = 0.25 \times \left[v(\hat{\theta}_1) + v(\hat{\theta}_2) \right]$

The second method estimates the mean of the two years using the combined data with the new weights described in Table 4. The mean estimates $\theta_w = (N_2\theta_2 + N_1\theta_1)/(N_2 + N_1)$, where N_1 and N_2 are the population sizes in the two surveys for the estimation group, which makes the mean weighted. When the population sizes in the two surveys are constant, the weighted mean reduces to the unweighted mean (i.e., $\hat{\theta}_m$). Over a short period of time, the population size of most groups would change very little. However, there may be subgroups increasing or decreasing in size rapidly by immigration, such as some of California's racial/ethnic groups. One advantage of using the combined data set

with the new weights is that it takes into account change in population size. This is because this method implicitly computes the population sizes by applying the final weights from multiple years simultaneously. Note that the weights are created to control for the population sizes of all racial/ethnic groups for the corresponding years in which the CHIS was conducted. Table 8 also shows the results applying the second method using the combined data (heading "Population weighted"); see example codes in the Appendix Section G.

Table 8 indicates that the two methods give very similar results for all but the smallest racial/ethnic group, i.e., Native Hawaiian & Pacific Islander group. The reason for this difference could be that the Native Hawaiian & Pacific Islander women ages 40 and older might have experienced rapid growth in its population.

Table 8. Mean Mammogram Usage for Women 40+ from CHIS 2001 and 2003

Characteristics	Mean		Population weighted	
	%	**SE**	**%**	**SE**
All	75.73	0.30	75.75	0.30
Age				
40-49	66.91	0.57	66.92	0.57
50-64	82.35	0.45	82.34	0.45
≥ 65	78.57	0.59	78.57	0.59
Race/Ethnicity				
Latino	69.71	0.91	69.72	0.92
Native Hawaiian & Pacific Islander	70.86	5.42	72.20	5.55
American Indian & Alaska Native	69.52	3.12	69.51	3.14
Asian	70.99	1.09	71.13	1.10
African American	77.35	1.18	77.33	1.19
White	78.32	0.36	78.32	0.36
Other/Multiple race	73.89	1.94	74.13	1.92
Metropolitan status				
Urban	75.83	0.31	75.84	0.32
Rural	73.45	0.80	73.41	0.81
Education				
<High school	68.92	0.93	68.92	0.93
High school	75.38	0.60	75.40	0.60
Some college +	78.34	0.36	78.34	0.36

Recall that the difference across two years for the Native Hawaiian & Pacific Islander subgroup in Table 3 was large at 23.43 percent. While combined data will provide larger sample sizes, caution should be used when reporting the average for this group in Table 8. As described at the beginning of this section, reporting such estimates may not be a good idea, as the average may represent neither of the years.

One appeal of the combined data is that it provides larger sample sizes. Hence, the stability of the estimates and the power of the analysis increase. Some organizations do not publish unstable estimates (often measured by their standard errors or coefficients of variation (*CV*)). Klein et al. (2002) provide a summary of data suppression criteria that

have been implemented for the Healthy People 2010 data system. The general rule in Klein et al. is to suppress estimates with a *CV* greater than 30%. The UCLA Center for Health Policy Research concurs with this rule and strongly recommends against using estimates with a *CV* greater than 30% (additional detail below). *CV* is a relative measure of the estimate's variability and calculated as $CV(\hat{\theta}) = se(\hat{\theta})/\hat{\theta}$, where $\hat{\theta}$ is an estimate for the parameter of interest θ and $se(\hat{\theta})$ is its standard error. *CV* is also known as the relative standard error (RSE), because it is the ratio of the standard error to the mean.

CV is a useful tool for determining the stability of estimates regardless of their magnitudes. However, one should be cautious in using *CV*s, as Kish (1965) points out. First, means or changes over time that are close to zero are likely to have very large *CV*s. Additionally, when there is no change over time, the *CV* of this change is infinity. Second, the standard error is the same for both categories of a binary variable, while their *CV*s differ once the estimates deviate from 0.5. For instance, the proportion of people who reported "yes" to some variable is 0.9 (therefore, "no" is 0.1) and $se(\text{yes}) = se(\text{no}) = 0.05$, which results in $CV(\text{yes}) = 5.5\%$ but $CV(\text{no}) = 50\%$. While presenting either "yes" or "no" will provide the same information, the fact that the cell-base *CV* for "yes" is small will indicate that the information about this variable is stable. A similar problem may be encountered in analyzing multinomial variables. The UCLA Center for Health Policy Research recommends using the following rule for calculating *CV*s of all categorical variables:

$$CV(\hat{\theta}) = se(\hat{\theta})/\hat{\theta}, \text{ if } \hat{\theta} \le 0.5 \text{ ; and}$$

$$CV(\hat{\theta}) = se(\hat{\theta})/(1-\hat{\theta}), \text{ otherwise.}$$

For categorical variables, therefore, the CV is calculated using the category with the smallest value as the denominator. This is a conservative method, because it minimizes possibilities of erroneously using estimates with low stability.

Table 9 shows mammogram usage for American Indian and Alaska Native women ages 40 to 45 who reside in urban areas by education, along with the *CV*s. The estimates for this group of women are 53.81% and 50.85% in respective years and 52.17% when combined. The *CV* decreases from 24% and 22% to 15% by using the combined data. The significant benefit of using combined data can be found for those with high school education as the *CV* of the estimate for the group decreases from 58% and 32% to 29%. Applying the criterion described above, this estimate becomes statistically stable when using the combined data file. However, in spite of the increased stability, one should be cautious when the estimates from respective years are different, because such estimates may represent neither of the combined years.

Table 9. Mammogram Usage for Urban American Indian and Alaska Native Women (Ages 40-45) from CHIS 2001 and 2003

Characteristics	2001		2003		Population weighted average	
	%	CV	%	CV	%	CV
All	53.81	0.24	50.85	0.22	52.17	0.15
Education						
<High school	30.90	1.44	14.10	1.63	24.29	0.81
High school	49.93	0.58	55.14	0.32	52.80	0.29
Some college +	68.51	0.18	57.89	0.23	62.01	0.15

Note: Estimates on this table are for illustration purposes only. Do not cite the estimates.

Goal 4. Estimating population totals using two years of data

Estimates of population totals using the combined data and their interpretation require a clear conceptualization. The combined data using the method described in this document produce population total estimates that sum weighted totals from all years. For instance, the weighted total for women who had a mammogram in these past two years is 10,962,660. This number indicates that there were a total of 10,962,660 women who had a mammogram in the last two cycles (or in the years of 2001 and 2003 but *not* 2001 through 2003) in California. As discussed, its interpretation may be difficult and confusing, and its utility is unclear. The UCLA Center for Health Policy Research recommends using yearly average totals, which can be calculated in the following manner:

$$Yearly\ Average\ Total = \sum_{s=1}^{S} \hat{T}_s \Big/ S \,,$$

where $\hat{T}_s$ is the population total estimate from year s with $s = 1,...,S$ and S is the number of years whose data are combined. Therefore, the yearly average total between 2001 and 2003 for women who had a mammogram in the past two years in California is 5,481,330 (=10,962,660/2).

References

Kish, L. (1965). *Survey Sampling*. New York: John Wiley & Sons.

Klein, R.J., Proctor, S.E., Boudreault, M.A., and Turczyn, K.M. (2002). "Healthy People 2010 Criteria for Data Suppression." *Healthy People 2010 Statistical Notes*. 24:1-12. Hyattsville, MD: National Center for Health Statistics.

Korn, E.L., and Graubard, B.I. (1999). *Analysis of Health Surveys*. New York: John Wiley & Sons.

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.

Westat. (2005). Combining CHIS 2001 and 2003 data files. Unpublished internal memo.

Appendix

SAS/SUDAAN and Stata Code Examples

The SAS-callable SUDAAN examples below work with SUDAAN 9.1 and later. Substitute `SUBGROUP` and `LEVELS` statements for `CLASS` statements if you use an earlier version. The Stata examples below work with Stata Version 9 and higher. For Stata Version 8, use `SVR` add-on module. (This document does not provide examples on specific add-on modules; however, this add-on application is freely available on the Web for downloading.)

A. Combined data file uses the following variables:

`RAKEDW0:` Full sample weight (included in both CHIS 2001 and CHIS 2003 files)

`RAKEWD1-RAKEDW80:` Replicate weights (included in both CHIS 2001 and CHIS 2003 files)

`MAM2YR:` Mammogram in the last 2 years (recoded)
  1 = Had mammogram in the last 2 years
  0 = Did not have mammogram in the last 2 years

`AGEA:` Age (recode)
  1 = 40-49
  2 = 50-64
  3 = 65+

`RACEHPR:` Race--UCLA CHPR definition (recoded)
  1 = Latino
  2 = Native Hawaiian & Pacific Islander
  3 = American Indian & Alaska Native
  4 = Asian
  5 = African American
  6 = White
  7 = Other/Multiple race

`UR_OMB:` Rural and urban--OMB (included in both CHIS 2001 and CHIS 2003 files)
  1 = Metropolitan
  2 = Non-metropolitan

`EDUC:` Education (recoded)
  1 = < High school
  2 = High school
  3 = Some college +

B. Creating a combined data file:

1. Combining CHIS 2001 and 2003
This example shows how to pool CHIS 2001 and 2003 data files to create a combined data file. For further analyses, the data file is restricted to females ages 40 and older.

SAS/SUDAAN:

```
data combined;
 set libname.chis 2001 data
      (in=in2001 keep=rakedw0 rakedw1-rakedw80 srage srsex aheduc ad14
      ad17 racehpra ur_omb where=((srage ge 40) & srsex=2))
      libname.chis 2003 data
      (in=in2003 keep=rakedw0 rakedw1-rakedw80 srage_p srsex aheduc
      ad14 ad17 racehpr ur_omb where=((srage_p ge 40) & srsex=2));

  if      in2001 then year=2001;
  else if in2003 then year=2003;

  if year=2001 then do;
      if ad17 in (1,2) then mam2yr=1;
      else if ad17 in (3,4,5)or ad14 = 2 then mam2yr=0;
      else mam2yr = .;
      racehpr=racehpra;
  end;
  if year=2003 then do;
      if ad17 in (1,2) then mam2yr=1;
      else if ad17 in (3,4,5)  or ad14 = 2 then mam2yr=0;
      else mam2yr = .;
      srage=srage_p;
  end;

  if 40=<srage<=49 then agea=1;
      else if 50=<srage<=64 then agea=2;
      else if 65=<srage<=110 then agea=3;
      else agea=.;

  if aheduc in (91, 1, 2) then educ=1;
      else if aheduc=3 then educ=2;
      else if 4=<aheduc<=10 then educ=3;
      else educ=.;

  ***Create new weight variables;
  fnwgt0 = rakedw0;
  array a_origwgts[80] rakedw1-rakedw80;
  array a_newwgts[160] fnwgt1-fnwgt160;
  do i = 1 to 80;
    if year=2001 then do;
      a_newwgts[i]    = a_origwgts[i];
      a_newwgts[i+80] = rakedw0;
    end;
    else if year=2003 then do;
      a_newwgts[i]    = rakedw0;
      a_newwgts[i+80] = a_origwgts[i];
    end;
  end;
run;
```

Stata:

```
log using "folder location\data_step.log", replace

*********reset memory if the dataset is bigger than 80m********

set memory 200m
set more off
set matsize 1600
set scrollbufsize 100000

***CHIS 2001 Adult data****
use "folder location\CHIS 2001 data"
keep rakedw0-rakedw80 srsex srage ur_omb racehpra ad14 ad17 aheduc ah47
keep if srsex==2 & srage>=40

gen year=2001
gen fnwgt0=rakedw0
rename racehpra racehpr

recode srage (40/49=1 40-49) (50/64=2 50-64) (65/110=3 65+), gen(agea)
label(Age)
recode aheduc (91 1/2=1 "<HS Grad")(3=2 "HS Grad")(4/10=3 ">HS Grad"),
gen(educ)

gen mam2yr= 1 if ad17 == 1
replace mam2yr= 1 if ad17 == 2
replace mam2yr = 0 if ad17 == 3
replace mam2yr = 0 if ad17 == 4
replace mam2yr = 0 if ad17 == 5
replace mam2yr = 0 if ad14 == 2
label variable mam2yr "Recent Mammogram"
label define mam2yrf 1 "Mammogram in the past 2 yrs" 0 "No mammogram in
the past 2 yrs"
label values mam2yr mam2yrf

for new fnwgt1-fnwgt160: gen X=0

foreach i of numlist 1/80{
          local j=`i'-0
          replace fnwgt`i'=rakedw`j'
}

foreach i of numlist 81/160{
          replace fnwgt`i'=rakedw0
}

save adult01 , replace

***CHIS 2003 Adult data****

use "folder location\CHIS 2003 data"

keep rakedw1-rakedw80 rakedw0 srsex srage ur_omb racehpr ad14 ad17
aheduc
keep if srsex==2 & srage>=40
gen year=2003
gen fnwgt0=rakedw0

recode srage (40/49=1 40-49) (50/64=2 50-64) (65/110=3 65+), gen(agea)
label(Age)
recode aheduc (91 1/2=1 "<HS Grad")(3=2 "HS Grad")(4/10=3 ">HS Grad"),
gen(educ)
```

```
gen mam2yr= 1 if ad17 == 1
replace mam2yr= 1 if ad17 == 2
replace mam2yr = 0 if ad17 == 3
replace mam2yr = 0 if ad17 == 4
replace mam2yr = 0 if ad17 == 5
replace mam2yr = 0 if ad14 == 2
label variable mam2yr "Recent Mammogram"
label define mam2yrf 1 "Mammogram in the past 2 yrs" 0 "No mammogram in
the past 2 yrs"
label values mam2yr mam2yrf

for new fnwgt1-fnwgt160: gen X=0

foreach i of numlist 1/80{
        replace fnwgt`i'=rakedw0
}

foreach i of numlist 81/160{
        local j=`i'-80
        replace fnwgt`i'=rakedw`j'
}

append using adult01

save "folder location\combined.dta", replace
```

2. Combining CHIS 2001, 2003 and 2005

This example shows how to pool CHIS 2001, 2003 and 2005 data files to create a combined data file. For further analyses, the data file is restricted to females ages 40 and older.

<u>**SAS/SUDAAN:**</u>

```
data combined;
 set libname.chis 2001 data
     (in=in2001 keep=rakedw0 rakedw1-rakedw80 srage srsex aheduc ad14
     ad17 racehpra ur_omb where=((srage ge 40) & srsex=2))
     libname.chis 2003 data
     (in=in2003 keep=rakedw0 rakedw1-rakedw80 srage_p srsex aheduc
     ad14 ad17 racehpr ur_omb where=((srage_p ge 40) & srsex=2))
     libname.chis 2005 data
     (in=in2005 keep=rakedw0 rakedw1-rakedw80 srage_p srsex aheduc
     ad14 ad17 racehpr ur_omb where=((srage_p ge 40) & srsex=2));

  if     in2001 then year=2001;
  else if in2003 then year=2003;
  else if in2005 then year=2005;

  if year=2001 then do;
     if ad17 in (1,2) then mam2yr=1;
     else if ad17 in (3,4,5)or ad14 = 2 then mam2yr=0;
     else mam2yr = .;
     racehpr=racehpra;
  end;

  if year=2003 then do;
     if ad17 in (1,2) then mam2yr=1;
```

```
      else if ad17 in (3,4,5)  or ad14 = 2 then mam2yr=0;
      else mam2yr = .;
      srage=srage_p;
  end;

  if year=2005 then do;
      if ad17 in (1,2) then mam2yr=1;
      else if ad17 in (3,4,5)  or ad14 = 2 then mam2yr=0;
      else mam2yr = .;
      srage=srage_p;
  end;

  if 40=<srage<=49 then agea=1;
      else if 50=<srage<=64 then agea=2;
      else if 65=<srage<=110 then agea=3;
      else agea=.;

  if aheduc in (91, 1, 2) then educ=1;
      else if aheduc=3 then educ=2;
      else if 4=<aheduc<=10 then educ=3;
      else educ=.;

  ***Create new weight variables;
  fnwgt0 = rakedw0;
  array a_origwgts[80] rakedw1-rakedw80;
  array a_newwgts[240] fnwgt1-fnwgt240;

  do i = 1 to 80;
    if year=2001 then do;
      a_newwgts[i]     = a_origwgts[i];
      a_newwgts[i+80]  = rakedw0;
      a_newwgts[i+160] = rakedw0;
    end;
    else if year=2003 then do;
      a_newwgts[i]     = rakedw0;
      a_newwgts[i+80]  = a_origwgts[i];
      a_newwgts[i+160] = rakedw0;
    end;
    else if year=2005 then do;
      a_newwgts[i]     = rakedw0;
      a_newwgts[i+80]  = rakedw0;
      a_newwgts[i+160] = a_origwgts[i];
    end;
  end;

run;
```

<u>Stata:</u>

```
log using "folder location\data_step.log", replace

*********reset memory if the dataset is bigger than 80m********

set memory 200m
set more off
set matsize 1600
set scrollbufsize 100000
```

```
***CHIS 2001 Adult data****
use "folder location\CHIS 2001 data"

keep rakedw0-rakedw80 srsex srage ur_omb racehpra ad14 ad17 aheduc
keep if srsex==2 & srage>=40

gen year=2001
gen fnwgt0=rakedw0
rename racehpra racehpr

recode srage (40/49=1 40-49) (50/64=2 50-64) (65/110=3 65+), gen(agea)
label(Age)
recode aheduc (91 1/2=1 "<HS Grad")(3=2 "HS Grad")(4/10=3 ">HS Grad"),
gen(educ)

gen mam2yr= 1 if ad17 == 1
replace mam2yr= 1 if ad17 == 2
replace mam2yr = 0 if ad17 == 3
replace mam2yr = 0 if ad17 == 4
replace mam2yr = 0 if ad17 == 5
replace mam2yr = 0 if ad14 == 2
label variable mam2yr "Recent Mammogram"
label define mam2yrf 1 "Mammogram in the past 2 yrs" 0 "No mammogram in
the past 2 yrs"
label values mam2yr mam2yrf

for new fnwgt1-fnwgt240: gen X=0

foreach i of numlist 1/80{
        local j=`i'-0
        replace fnwgt`i'=rakedw`j'
}

foreach i of numlist 81/240{
        replace fnwgt`i'=rakedw0

}

save adult01 , replace

***CHIS 2003 Adult data****

use "folder location\CHIS 2003 data"

keep rakedw1-rakedw80 rakedw0 srsex srage ur_omb racehpr ad14 ad17
aheduc
keep if srsex==2 & srage>=40
gen year=2003
gen fnwgt0=rakedw0

recode srage (40/49=1 40-49) (50/64=2 50-64) (65/110=3 65+), gen(agea)
label(Age)
recode aheduc (91 1/2=1 "<HS Grad")(3=2 "HS Grad")(4/10=3 ">HS Grad"),
gen(educ)

gen mam2yr= 1 if ad17 == 1
replace mam2yr= 1 if ad17 == 2
replace mam2yr = 0 if ad17 == 3
replace mam2yr = 0 if ad17 == 4
replace mam2yr = 0 if ad17 == 5
replace mam2yr = 0 if ad14 == 2
label variable mam2yr "Recent Mammogram"
label define mam2yrf 1 "Mammogram in the past 2 yrs" 0 "No mammogram in
```

```
the past 2 yrs"
label values mam2yr mam2yrf

for new fnwgt1-fnwgt240: gen X=0

foreach i of numlist 1/80{
        replace fnwgt`i'=rakedw0
}

foreach i of numlist 81/160{
        local j=`i'-80
        replace fnwgt`i'=rakedw`j'
}
foreach i of numlist 161/240{
        replace fnwgt`i'=rakedw0
}


append using adult01
save adult0103 , replace


***CHIS 2005 Adult data****

use "folder location\CHIS 2005 data"

keep rakedw1-rakedw80 rakedw0 srsex srage ur_omb racehpr ad14 ad17
aheduc
keep if srsex==2 & srage>=40
gen year=2005
gen fnwgt0=rakedw0

recode srage (40/49=1 40-49) (50/64=2 50-64) (65/110=3 65+), gen(agea)
label(Age)
recode aheduc (91 1/2=1 "<HS Grad")(3=2 "HS Grad")(4/10=3 ">HS Grad"),
gen(educ)

gen mam2yr= 1 if ad17 == 1
replace mam2yr= 1 if ad17 == 2
replace mam2yr = 0 if ad17 == 3
replace mam2yr = 0 if ad17 == 4
replace mam2yr = 0 if ad17 == 5
replace mam2yr = 0 if ad14 == 2
label variable mam2yr "Recent Mammogram"
label define mam2yrf 1 "Mammogram in the past 2 yrs" 0 "No mammogram in
the past 2 yrs"
label values mam2yr mam2yrf

for new fnwgt1-fnwgt240: gen X=0

foreach i of numlist 1/160{
        replace fnwgt`i'=rakedw0
}

foreach i of numlist 161/240{
        local j=`i'-160
        replace fnwgt`i'=rakedw`j'
}

append using adult0103

save "folder location\combined.dta", replace
```

C. Design statements when using the combined data file:

Design statements can be used in various SUDAAN procedures to analyze the combined dataset. The Stata design statement (SVYSET) is needed before the analysis. It does not need to be repeated for each analysis in Stata, once applied.

SAS/SUDAAN:

```
PROC PROCEDURENAME data=combined design=jackknife;
  weight fnwgt0;
  jackwgts fnwgt1-fnwgt160 / adjjack=1;
```

Stata:

```
use "folder location\combined.dta"

svyset [pw=fnwgt0], jkrw(fnwgt1-fnwgt160, multiplier(1)) vce (jack) mse
```

D. Estimating changes using the combined data file:

This example shows how to use SUDAAN's PROC DESCRIPT and Stata's SVY MEAN and LINCOM with a combined data file to estimate changes in mammogram usage from CHIS 2001 to CHIS 2003 by various subgroups. This code produces the results shown in Table 3. The same results can be achieved by analyzing the 2001 and 2003 data sets separately, then performing the calculations summarized in Table 2.

SAS/SUDAAN:

```
PROC DESCRIPT data=combined design=jackknife;
weight fnwgt0;
jackwgts fnwgt1-fnwgt160 / adjjack=1;
class agea racehpr ur_omb educ year/nofreq;
var mam2yr;
catlevel 1;
diffvar year=(2003 2001) / name="Change from 2001 to 2003";
tables  agea racehpr ur_omb educ;
print nsum percent lowpct uppct/style=nchs;
run;
```

Stata:

```
svy: mean mam2yr, over (year)
lincom [mam2yr]2003 - [mam2yr]2001
```

For Stata, it is necessary to recode each category of the variable as a new variable before obtaining the mean percentage of mammogram usage and the difference between two years.

```
generate age1= (agea ==1)
```

```
svy, subpop(age1): mean mam2yr, over (year)
lincom [mam2yr]2003 - [mam2yr]2001

generate age2= (agea ==2)
svy, subpop(age2): mean mam2yr, over (year)
lincom [mam2yr]2003 - [mam2yr]2001

generate age3= (agea ==3)
svy, subpop(age3): mean mam2yr, over (year)
lincom [mam2yr]2003 - [mam2yr]2001


generate race1= (racehpr ==1)
svy, subpop(race1): mean mam2yr, over (year)
lincom [mam2yr]2003 - [mam2yr]2001

generate race2= (racehpr ==2)
svy, subpop(race2): mean mam2yr, over (year)
lincom [mam2yr]2003 - [mam2yr]2001

generate race3= (racehpr ==3)
svy, subpop(race3): mean mam2yr, over (year)
lincom [mam2yr]2003 - [mam2yr]2001
generate race4= (racehpr ==4)
svy, subpop(race4): mean mam2yr, over (year)
lincom [mam2yr]2003 - [mam2yr]2001

generate race5= (racehpr ==5)
svy, subpop(race5): mean mam2yr, over (year)
lincom [mam2yr]2003 - [mam2yr]2001

generate race6= (racehpr ==6)
svy, subpop(race6): mean mam2yr, over (year)
lincom [mam2yr]2003 - [mam2yr]2001

generate race7= (racehpr ==7)
svy, subpop(race7): mean mam2yr, over (year)
lincom [mam2yr]2003 - [mam2yr]2001


generate ur_omb1= (ur_omb ==1)
svy, subpop(ur_omb1): mean mam2yr, over (year)
lincom [mam2yr]2003 - [mam2yr]2001

generate ur_omb2= (ur_omb ==2)
svy, subpop(ur_omb2): mean mam2yr, over (year)
lincom [mam2yr]2003 - [mam2yr]2001


generate educ1= (educ ==1)
svy, subpop(educ1): mean mam2yr, over (year)
lincom [mam2yr]2003 - [mam2yr]2001

generate educ2= (educ ==2)
svy, subpop(educ2): mean mam2yr, over (year)
lincom [mam2yr]2003 - [mam2yr]2001
```

```
generate educ3= (educ ==3)
svy, subpop(educ3): mean mam2yr, over (year)
lincom [mam2yr]2003 - [mam2yr]2001
```

E. Estimating changes, controlling for other factors, using the combined data file:

This example shows how to use SUDAAN's PROC RLOGIST and Stata's SVY LOGIT with a combined data file to investigate changes in mammogram usage from CHIS 2001 to CHIS 2003, controlling for other factors. The data set *pacific* was formed by restricting the data set *combined* to Native Hawaiian & Pacific Islander women ages 40 and older. The data set *asian* was formed by restricting *combined* to Asian women ages 40 and older. This code produces the results shown in Table 5.

<u>**SAS/SUDAAN:**</u>

```
data pacific; set combined; if racehpr=2; run;

PROC RLOGIST data=pacific filetype=sas design=jackknife;
weight fnwgt0;
jackwgts fnwgt1-fnwgt160 / adjjack=1;
class agea ur_omb educ year/nofreq;
model mam2yr = agea ur_omb educ year;
reflevel agea=1 ur_omb=1 educ=1 year=2001;
run;

data asian; set combined; if racehpr=4; run;

PROC RLOGIST data=asian filetype=sas design=jackknife;
weight fnwgt0;
jackwgts fnwgt1-fnwgt160 / adjjack=1;
class agea ur_omb educ year/nofreq;
model mam2yr = agea ur_omb educ year;
reflevel agea=1 ur_omb=1 educ=1 year=2001;
run;
```

<u>**Stata:**</u>

```
generate pacific= (racehpr==2)
xi: svy, subpop(pacific): logit mam2yr i.agea i.ur_omb i.educ i.year

generate asian= (racehpr==4)
xi: svy, subpop(asian): logit mam2yr i.agea i.ur_omb i.educ i.year
```

F. Estimating changes, including an interaction between race/ethnicity and survey year, controlling for other factors, using the combined data file:

This example shows another way to use SUDAAN's PROC RLOGIST and Stata's SVY LOGIT with a combined data file to investigate change in mammogram usage from CHIS

2001 to CHIS 2003, controlling for various other factors but including an interaction between race/ethnicity and year. This code produces the results shown in Table 6.

G. Estimating means for multiple years using replicate weights in the combined data file:

This example shows how to use SUDAAN's PROC DESCRIPT and Stata's SVY MEAN with a combined data file to estimate the mean percentage in mammogram usage for both years by various subgroups using replicate weights. This code produces the results shown in Table 8.