November 4, 2009

Variance Estimation and Other Analytic Issues in the 1997-2005 NHIS
(Adapted from Appendices III and VII of the 2005 NHIS Survey Description Document,
and Appendix III of the 2006 NHIS Survey Description Document)

Introduction

The data collected in the NHIS are obtained through a complex, multistage sample design that involves stratification, clustering, and oversampling of specific population subgroups. The final weights provided for analytic purposes have been adjusted in several ways to yield valid estimates for the civilian, noninstitutionalized population of the United States. As with any variance estimation methodology, the techniques presented here involve several simplifying assumptions about the design and weighting scheme applied to the data. The first part of this document provides guidelines for data users based on simplified concepts of the NHIS sample design structure so that users may compute reasonably accurate standard error estimates. The second part of this document provides guidance for merging NHIS data files, and combining NHIS data across years.

There are several available software packages for analyzing complex samples. The Web site *Summary of Survey Analysis Software,* currently located at:

http://www.fas.harvard.edu/~stats/survey-soft/survey-soft.html ,

provides references for and a comparison of different software alternatives for the analysis of complex data. Analysts at NCHS generally use the software package SUDAAN® (Shah et al. 1997) to produce accurate standard error estimates. In this document, examples of SUDAAN computer code are provided for illustrative purposes. Examples also are provided for the Stata, SPSS, SAS, R, and VPLX software packages. However, the appropriate application of these procedures is the ultimate responsibility of data users, and the example command code is *not* "guaranteed." Both the computer command code and methods are subject to change without notification to the user. NCHS strongly recommends that NHIS data be analyzed under the direction of or in consultation with a statistician who is cognizant of sampling methodologies and techniques for the analysis of complex survey data.

CAUTION. Users are reminded that the use of standard statistical procedures that are based on the assumption that data are generated via simple random sampling (SRS) generally will produce incorrect estimates of variances and standard errors when used to analyze data from the NHIS. The clustering protocols that are used in the multistage selection of the NHIS sample require other analytic procedures, as described below. Analysts who apply SRS techniques to NHIS data generally will produce standard error estimates that are, on average, too small, and are likely to produce results that are subject to excessive Type I error.

Conceptual NHIS design for 1995-2005

Thorough discussions of the NHIS design, the methods used for weighting data, and the

methods used for variance estimation are beyond the scope of this document but are provided elsewhere (NCHS 1999; NCHS 2000). This document outlines the basic ideas published in these technical reports (NCHS 1999; NCHS 2000).

To achieve sampling efficiency and to keep survey operations manageable, cost-effective, and timely, the NHIS survey planners used multistage sampling techniques to select the sample of persons and households for the NHIS. These multistage methods partition the target universe into several nested levels of strata and clusters. The NHIS target universe is defined as all dwelling units in the U.S. that contain members of the civilian noninstitutionalized population. As the NHIS is conducted in a face-to-face interview format, a simple random sample of dwelling units would be too dispersed throughout the nation; as a result, the costs of interviewing a simple random sample of 40,000 households would be prohibitive. Also, specific population subgroups, such as black and/or Hispanic households, would not be sampled sufficiently under a simple random sample design. To achieve survey objectives subject to resource constraints, the NHIS uses methods of clustering, stratification, and oversampling of specific population subgroups.

First, the target universe was partitioned into approximately 1,900 primary sampling units (PSUs), which are single counties, groups of adjacent counties (or equivalent jurisdictions), or metropolitan areas. These PSUs vary in population size and number of jurisdictions. Cost-effective field operations and efficient sampling result in those PSUs with the largest populations (e.g., the New York City metropolitan area) being sampled with certainty, and the smaller universe PSUs being represented by a sample. These smaller PSUs are called non-self-representing (NSR) or non-certainty PSUs. The universe of NSR PSUs is stratified using multiple criteria consistent with NHIS objectives. The NSR PSUs were stratified first at the state level according to metropolitan status (metro or non-metro). If a particular NSR stratum in a given state contained a large population, then it was further stratified by aggregate-level poverty rates. Thus, the number of NSR strata varies from state to state, and the number of PSUs varies from stratum to stratum. Once these strata were defined, a sample of PSUs was selected; within most NSR strata, two PSUs were selected without replacement with probability proportional to population size, and the SR PSUs were selected with certainty. Within a few NSR strata with smaller population sizes, only one PSU was drawn.

The U.S. Bureau of the Census partitioned each selected NSR or SR PSU into substrata of Census blocks or combined blocks based on the concentrations of black and Hispanic populations. These race and ethnicity density substrata were defined according to the population concentrations from the 1990 Decennial Census. New housing within a PSU was included as its own substratum in order to produce the most current sample of households. Each PSU could be partitioned into up to 21 substrata of dwelling units. Large metropolitan SR PSUs tend to have many substrata, while the NSR PSUs tend to have only a few.

Sampling within the PSU substrata is complex and involves clustering dwelling units within each substratum. These clusters form a universe of Secondary Sampling Units (SSUs). A systematic sample of SSUs is selected to represent each substratum. Each race and ethnicity density substratum has its own sampling rate for SSU selection.

Within each selected SSU all households containing black or Hispanic persons are

selected for interview, while only a sample of other households is selected. The non-black, non-Hispanic households are sampled at different rates within the 21 substrata. For selected households, the NHIS collects some information on all household members, and additional information is obtained for randomly selected persons in each household. For example, one adult per family is randomly selected for interview with the sample adult questionnaire.

This hierarchy of sampling allows the creation of household- and person-level base weights. Each base weight is the product of the inverses of the probability of selection at each sampling stage. Roughly speaking, the base weight is the number of population units a sampled unit represents. Under ideal sampling conditions, and if 100% response occurred, a base-weighted sample total will be an unbiased estimator for the true total in the target population. In practice, however, the base weights are adjusted for non-response, and ratio-adjusted to create final sampling weights. The final person-level weights are adjusted according to a quarterly poststratification by 88 age/sex/race/ethnicity classes based on population estimates produced by the U.S. Bureau of the Census. Most other weights receive some form of ratio adjustment as well.

Internally, NCHS uses the design and weighting information to formulate appropriate variance estimates for NHIS statistics. While recognizing the need to provide accurate information, NCHS also must adhere to the Public Health Service Act (Section 308(d)) that forbids the disclosure of any information that may compromise the confidentiality promised to its survey respondents. Consequently, much of the NHIS design information cannot be publicly released, and other data are either suppressed or recoded to insure confidentiality. In order to satisfy this disclosure constraint, many of the original design strata, substrata, PSUs, and SSUs are masked for public release by applying techniques to cluster, collapse, mix, and partition the original design variables. Through this process the original NHIS design variables are transformed into public use design variables (i.e., STRATUM and PSU). The public use design structures perform reasonably well when compared to internal NCHS design structures (NCHS 2000). The sampling weights have not been changed in any way for the public data. Data users who want access to internal NCHS data have the option of accessing data through the NCHS Research Data Center. For further information, refer to **http://www.cdc.gov/nchs/r&d/rdc.htm**.

Design Information Available on the NHIS Public Use Data

The 1997-2005 Household and Person public use files contain the design variables necessary for variance estimation; Table 1 provides a summary of the Person file variables. Users should check the Variable Layouts of the Household and Person files for any additional information about these variables. Note that for the 2004 NHIS data, the design variables (STRATUM and PSU) are not included on the Family, Sample Child and Sample Adult files. For the 2005 NHIS data, the design variables are included on the Sample Child and Sample Adult files, but not the Family file.

Table 1. Variables Used for Variance Estimation, 1997-2005 NHIS Person File

Variable Name	Variable Label
STRATUM	Stratum for variance estimation
PSU	PSU for variance estimation
WTFA	Weight - Final, annual Person weight

As discussed above, in order to mask true geographical locations, the STRATUM and PSU levels are pseudo-levels or simplified versions of the true NHIS sample design variables. Analysts are cautioned that these simplified design structures do not support geographical analyses below the Census Region level.

CAUTION. Significant changes were made to the Stratum and PSU values beginning with the 1997 survey year. More strata have been provided (compared to the 1995 public release) to improve statistical efficiency in various statistical estimation procedures. The sample design variables provided on the 1997-2005 NHIS public use data files are *not* comparable to those of 1995-1996. Users are cautioned that variance estimation structures discussed here are for individual survey years only, not for pooled analyses of multiple years of the NHIS. Refer to the section "Variance Estimation for Pooled Analyses of Adjacent Years of the NHIS" in this document for variance estimation guidance for pooled analyses of adjacent years of the NHIS.

Variance Estimation Method for Public Use Data

The method described below is applicable to all 1997-2005 NHIS public use data, except the Injury/Poisoning Episode file and the Verbatim Injury/Poisoning Episode file. Refer to the end of this section for some supplemental guidance regarding the 2003 and 2005 Sample Child files.

For this method of variance estimation, the NHIS sample is treated as having 339 strata, each containing two sampled PSUs. While in reality no PSU was sampled more than once, the limited public release design information requires a mathematical simplification that the PSUs be treated as if they were sampled with replacement (WR). This public use method provides slightly more conservative (larger) standard errors than the variance estimation method that is applied internally by analysts at NCHS (NCHS 2000). Additionally, this public use method is applicable in many statistical packages for complex survey data that require exactly two sample PSUs per stratum. Moreover, this method is robust when analyzing subsetted or subgroup data (see the section "Subsetted Data Analysis" below).

When implementing this public use method, users should observe 678 PSUs when analyzing the full database. The simplified design structure can be specified with the following statements in SUDAAN:

> *PROC* <DESCRIPT, CROSSTAB, ...>**...**   **DESIGN = WR ;**
> *NEST* **STRATUM PSU ;**
> *WEIGHT*   **WTFA ;**

Note that SUDAAN requires that the input file be sorted by the variables listed on the NEST

statement (i.e., STRATUM and PSU). Design statements for other data files should use the appropriate weight variables found on these files.

Corresponding statements for other software packages are as follows:

Stata svy:

SVYSET **[PWEIGHT=WTFA]**,*STRATA*(**STRATUM**)*PSU*(**PSU**)
SVY: MEAN <name of variable to be analyzed for average>
Or
SVY: PROPORTION <name of variable to be analyzed for percentage/proportion>

SPSS csdescriptives (for averages) or cstabulate (for percentages/proportions):

One needs first to define a "plan file" with information about the weight and variance estimation, e.g.:

CSPLAN ANALYSIS
/PLAN FILE="< file name >*"*
/PLANVARS ANALYSISWEIGHT=**WTFA**
/DESIGN STRATA=**STRATUM** *CLUSTER*=**PSU**
/ESTIMATOR TYPE=**WR**.

And then refer to the plan file when using csdescriptives or cstabulate, e.g.:

CSDESCRIPTIVES
/PLAN FILE="< file name >*"*
/SUMMARY VARIABLES =<name of variable to be analyzed>
/MEAN.

CSTABULATE
/PLAN FILE="< file name >*"*
/TABLES VARIABLES =<name of variable to be analyzed>
/CELLS TABLEPCT.

SAS proc surveymeans (for averages) or surveyfreq (for percentages/proportions) :

PROC SURVEYMEANS;
STRATA **STRATUM**;
CLUSTER **PSU**;
WEIGHT **WTFA**;
VAR <name of variable to be analyzed>;
RUN;

PROC SURVEYFREQ;
STRATA **STRATUM**;

CLUSTER **PSU**;
WEIGHT **WTFA**;
TABLES <name of variable to be analyzed>;
RUN

R (including the "survey" package):
(note: R syntax is case-sensitive)

 # load survey package
require(survey)
 # create data frame with NHIS design information, using existing data frame of NHIS data
*nhissvy <- svydesign(id=~***psu***, strata=~***stratum***,*
                            nest = TRUE,
                            *weights=~***wtfa***,*
                            data=< existing data frame name>)
svymean(~<name of variable to be analyzed>,design=nhissvy)

note: svymean will produce proportions for "factor variables". Consult the R documentation (http://cran.r-project.org/manuals.html) for details.

VPLX:

In the CREATE step, include the following statements:

STRATUM   **STRATUM**
CLUSTER   **PSU**
WEIGHT   **WTFA**

Then specify the variable to be analyzed in the DISPLAY step:

LIST   *MEAN*(<name of variable to be analyzed>)

VPLX can produce percentages by including a CAT statement in the CREATE step. Consult the VPLX documentation (http://www.census.gov/sdms/www/vdoc.html) for details.

CAUTION. A rule of thumb to calculate the number of degrees of freedom to associate with a standard error is the quantity *number of PSUs - number of strata*. Typically, this rule is applied to a design with two PSUs per stratum and when the variance components by stratum are roughly the same magnitude. The applicability of this rule depends upon the variable of interest and its interaction with the design structure (for additional information, see Chapter 5 of Korn and Graubard 1999). Given this rule of thumb, the number of degrees of freedom for the public use method described above is 339. This number of degrees of freedom is used to determine the *t*-statistic, its associated percentage points, p-values, standard error, and confidence intervals. As the number of degrees of freedom becomes large, the distribution of the *t*-statistic approaches the standard normal distribution. For example, with 120 degrees of freedom, the 97.5 percentage point of the t_{120} distribution is 1.980, while the 97.5 percentage point of the standard normal distribution

is 1.960. If a variable of interest is distributed across most of the NHIS PSUs, a normal distribution assumption may be adequate for analysis since the number of degrees of freedom would be large. The user should consult a mathematical statistician for further discussion.

Supplemental guidance for the 2003 and 2005 Sample Child files:

The 2003 Sample Child file does not have any records for PSU 2 in Stratum 185, and the 2005 Sample Child file does not have any records for PSU 2 in Stratum 185 or PSU 1 in Stratum 297. This situation is not handled consistently by all contemporary software alternatives for the analysis of complex data; some software (e.g., Stata 9) will return missing values for the standard error estimates, while other software (e.g., SPSS, SAS survey procedures) will return standard error estimates that are slightly smaller than they should be.

NCHS has created supplemental files with dummy records to fill in the missing sample design information for the 2003 and 2005 Sample Child files. 2003samchild.dat has one record, and 2005samchild.dat has two records. The layout of 2003samchild.dat is the same as the 2003 Sample Child file, and the layout of 2005samchild.dat is the same as the 2005 Sample Child file. The weight variables for the supplemental records, WTIA_SC and WTFA_SC, have been set equal to 1. Users who wish to use the dummy records for analyses can use the same programs for both the Sample Child file and the supplemental file, merge the two files into one, and then sort the concatenated file. The guidance provided below for Subsetted Data Analysis should then be used to select the records with WTIA_SC and/or WTFA_SC>1 for analysis. For example, for SUDAAN, use SUBPOPN WTFA_SC>1.

SUDAAN users do not have to include the supplemental files in their analyses. If the MISSUNIT option is used (see Strategy 2 below in the Subsetted Data Analysis section), the outcome is the same as using the supplemental files and the SUBPOPN statement.

Stata 10 users do not have to include the supplemental files in their analyses. The option singleunit(centered) in the svyset statement is equivalent to the MISSUNIT option in SUDAAN.

R users (including the "survey" package) do not have to include the supplemental files in their analyses. The statement:

options("survey.lonely.psu"="adjust")

is equivalent to the MISSUNIT option in SUDAAN.

Subsetted Data Analysis

Frequently, studies using NHIS data are restricted to specific population subgroups, e.g., persons aged 65 and older. Some users delete all records outside of the domain of interest (e.g., persons aged less than 65 years) in order to work with smaller data files and run computer jobs more quickly. This procedure of keeping only selected records (and list-wise deleting other records) is called subsetting the data. With a subsetted dataset that is appropriately weighted, correct point estimates (e.g., estimates of population subgroup means) can be produced. **However,**

most software packages that analyze complex survey data incorrectly compute standard errors for subsetted data. When complex survey data are subsetted, oftentimes the sample design structure is compromised because the complete design information is not available; subsetting data deletes important design information needed for variance estimation. Note that SUDAAN has a SUBPOPN option that allows the targeting of a subpopulation while using the full (unsubsetted) data file containing the design information for the entire sample. (See a SUDAAN manual for more information.) **NCHS recommends that subpopulation analyses be carried out using the full data file and the SUBPOPN option in SUDAAN, or an equivalent procedure with another complex design variance estimation software package.**

Strategy 1 (recommended) Use the SUBPOPN statement with the method described above for the full Person file dataset:

> *PROC ...* **DESIGN = WR ;**
> *NEST* **STRATUM PSU ;**
> *WEIGHT* **WTFA ;**
> *SUBGROUP* **(variable names);**
> *LEVELS ...* **;**
> *SUBPOPN* **RACRECI2=2 & SEX=2 / NAME="Analysis of African American women;"**

Using the full dataset with the SUBPOPN statement in this example would constrain this analysis to African American women only (RACRECI2= 2 for black and SEX = 2 for female). Use of the SUBPOPN statement is equivalent to subsetting the dataset, except that any resulting variance estimates are based on the full design structure for the complete dataset.

Strategy 2 (not recommended, except when Strategy 1 is infeasible) Use the MISSUNIT option on the NEST statement with the method described above for subsetted data:

> *NEST* **STRATUM PSU / MISSUNIT ;**

In a WR design with exactly two PSUs per stratum, when some PSUs are removed from the database through the listwise deletion of records outside the population of interest, the MISSUNIT option in SUDAAN "fixes" the estimation to avoid errors due to the presence of strata with only one PSU. However, in general there is no guarantee that the variance estimates obtained by this method are equivalent to those obtained using Strategy 1. Other calculations, such as those for design effects, degrees of freedom, standardization, etc., may need to be carried out differently. Users are responsible for verifying the correctness of their results based on subsetted data.

Implementing Strategy 1 in other software packages can be accomplished as follows:

Stata svy:

Add SUBPOP to the SVY statement, e.g.:

SVY,SUBPOP(**RACRECI2==2 & SEX==2**)*: MEAN* <name of variable to be analyzed>

SPSS csdescriptives or cstabulate:

One must first define an indicator variable, e.g.:

DO IF (RACRECI2 EQ 2 AND SEX EQ 2).
  COMPUTE SUBGRP=1.
ELSE.
  COMPUTE SUBGRP=0.
END IF.

And then refer to the indicator variable in csdescriptives or cstabulate, e.g.:

CSDESCRIPTIVES (or CSTABULATE)
/SUBPOP TABLE=SUBGRP

It is **very important** that the indicator variable is defined for all data records, otherwise an invalid result can occur.

SAS proc surveymeans or surveyfreq:

One must first define an indicator variable, e.g.:

IF RACRECI2=2 & SEX=2   *THEN SUBGRP=1;*
                          ELSE SUBGRP=0;

And then refer to the indicator variable in proc surveymeans using the DOMAIN statement, e.g.:

PROC SURVEYMEANS;
DOMAIN SUBGRP;

Proc surveyfreq does not have a DOMAIN statement. Instead, include the indicator variable in the TABLES specification:

PROC SURVEYFREQ;
TABLES SUBGRP<name of variable to be analyzed>;

As with SPSS, it is **very important** that the indicator variable is defined for all data records, otherwise an invalid result can occur.

R (including the "survey" package):

After applying the svydesign function to a data frame that contains the entire NHIS sample file being analyzed, create a new data frame using the criteria that define the subgroup of interest. Note that R is very "feisty" when testing for equality, hence the syntax that follows specifies the subgroup of interest without using an equality test.

 # subset for racreci2=2 & sex=2 without using equal signs
subgrp <- subset(nhissvy,((racreci2>1) & (racreci2<3) & (sex>1)))
svymean(~<name of variable to be analyzed>,design=subgrp)

VPLX:

In the CREATE step, define one or more CLASS variables that can be used to specify the criteria that define the subgroup of interest.

COPY RACRECI2 INTO RACECAT
COPY SEX INTO SEXCAT
CLASS RACECAT (1/2/3-HIGH)
CLASS SEXCAT (1/2)

The second category of RACECAT, crossed with the second category of SEXCAT, defines the subgroup of interest.

Then, specify the variable to be analyzed in the DISPLAY step, and specify the subgroup of interest as well:

LIST   MEAN(<name of variable to be analyzed>) */CLASS* RACECAT(2)*SEXCAT(2)

Note that the specification of RACECAT(2) and SEXCAT(2) is to the second category of each variable, which happens to be the value "2" in both cases in this example. Specification of RACECAT(3) would include all values of RACRECI2 of 3 and higher ("3-HIGH").

Variance Estimation for Pooled Analyses of Adjacent Years of the NHIS

Adjacent years of NHIS data sometimes are combined for a pooled analysis, e.g., 2004 and 2005, or 2002-2004. A pooled analysis might be done, for example, to increase the sample size for some small population. An estimate from a pooled analysis can be interpreted to be an estimate for the midpoint of the time interval of the pooled data.

The sampling weights for pooled data should be adjusted; otherwise, estimates of totals will be too high. For example, the estimated total U.S. civilian noninstitutionalized population from two years of pooled data, using unadjusted weights, would be about twice as large as it should be. A simple, valid weight adjustment procedure that NCHS recommends is to divide each sample weight in the pooled dataset by the number of years that are being pooled; e.g., divide by 2 when two years of data are combined, divide by 3 when three years of data are combined, etc. A sophisticated user may want to consider an alternative weight adjustment method that would minimize the variance of a particular estimate; however, in general, if the sample sizes are similar in the data years being combined, the simple procedure and the sophisticated alternative would give a similar adjustment.

Variance estimation for pooled analyses falls into one or more of the following three

10

classifications:

#1. The years being pooled fall within the same sample design period with the same public use design variables, and no changes were made to the design variables within the years being pooled.
#2. The years being pooled fall into different sample design periods (e.g., 1985-1994, 1995-2005).
#3. The years being pooled fall within the same sample design period, and there were changes to the public use design variables (e.g., from 1995-1996 to 1997-2005).

For #1, the sample has been drawn from the same geographic areas (same sample design), and the definitions of the variables used for public use variance estimation have not changed within the time period being analyzed. A valid method for variance estimation is to treat the pooled data like one year of data with a very large sample size. It is not correct to treat the different data years as being statistically independent, because the samples for the different years were drawn from the same geographic areas. Treating different data years as being statistically independent generally will lead to standard error estimates that are too small, and standard error estimates of contrasts (differences) between years would tend to be too large if the yearly estimates are positively correlated.

For #2, the different sample design periods should be treated as statistically independent. If there are multiple years of data being used for one or both design periods, each group should be treated in a similar manner as described in #1, assuming that the design variables within each group were unchanged. For example, if 1992-1995 NHIS data were pooled, the #1 procedure applies for the 1992-1994 data, and that aggregate is treated as being statistically independent from the 1995 data.

Note that it may be necessary to create new design variables to carry out this type of analysis. For example, consider an analysis of 1992-1995 NHIS data. The design variables have different names in the two sample design periods, and the stratum identifiers have different lengths. Referring to the first method described in "Variance Estimation for Person Data Using SUDAAN and the National Health Interview Survey (NHIS) Public-Use Person Data Files, 1987-94", currently available online at **http://www.cdc.gov/nchs/nhis/sudaan.htm**, the design variables for the 1992-1994 data are CSTRATUM (stratum), CPSU (PSU), and WTF (weight), while they are STRATUM, PSU, and WFTA, respectively, for the 1995 data. Suppose the names of the new design variables are NSTRATUM (stratum), NPSU (PSU), and NWT (weight). One method to create values for NSTRATUM that are of consistent length and take account of the different sample design periods is to do the following: for the 1992-1994 data, where the CSTRATUM values are 1, 2, ..., 62, first change these to 001, 002, ..., 062 (consistent length with STRATUM), and then do something to make them distinct from the STRATUM values, such as put a "1" in front: 1001, 1002, ..., 1062. For the 1995 data, where the STRATUM values are 1, 2, ..., 339, first change these to 001, 002, ..., 339, and then do something to make them distinct from the CSTRATUM values, such as put a "2" in front: 2001, 2002, ..., 2339. NPSU can be set equal to CPSU for the 1992-1994 data, and equal to PSU for the 1995 data, as both CPSU and PSU are of length one. NWT can be set equal to WTF/4 for the 1992-1994 data, and to WFTA/4 for the 1995 data.

For #3, no entirely satisfactory approach is available. Grouping of years should be done

over the periods where the same public use design variables are present (i.e., like #1). Then, for combining across years where there were changes to the public use design variables, the only option is to carry out an analysis as if the data years were statistically independent. For example, if 1995-1999 NHIS data were pooled, the #1 procedure applies for 1995-1996, and 1997-1999; then, the only alternative is to treat these two groups as statistically independent. The resulting standard error estimates may be too small, and standard error estimates of contrasts between years might be too large if the yearly estimates are positively correlated.

References

Cochran, W.G. (1977), *Sampling techniques* (3rd edition), John Wiley & Sons.

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

National Center for Health Statistics (1999), *National Health Interview Survey: Research for the 1995-2004 redesign*, Vital and Health Statistics, Series 2, No. 126.

National Center for Health Statistics (2000), *Design and Estimation for the National Health Interview Survey, 1995-2004*, Vital and Health Statistics, Series 2, No. 130.

Shah, B.V., Barnwell, B.G. and Bieler, G.S. (1997), *SUDAAN User's Manual; Release 7.5*, Research Triangle Institute, Research Triangle Park, NC.

Merging Data Files and Combining Years of Data in the NHIS

NHIS data files can be merged within years as well as combined across years. The purpose of merging data *within* a particular data year is to incorporate variables from different data files when respondents are common to both files, thereby increasing the number of variables available for analysis for a given individual. In contrast, the purpose behind combining NHIS data files *across* survey years is to combine respondents from different data years while retaining variables common to both files, thereby increasing the number of respondents (as long as the same variables are found in both files) and the precision of estimates.

Merging Data Files

Unlike survey years prior to 2004, variables are not generally repeated on multiple data files in the 2004 and 2005 NHIS. **As a result, users may find it necessary to perform additional merging of the 2004 and 2005 files in order to analyze the data.** Each data file contains household, family, and person record identifiers that make merging possible within the 2004 or 2005 files. Once the data files are sorted by record identifiers common to each file, merging is straightforward. Below is an example of a SAS program that will merge data files **within** an NHIS data year. Using the household, family, and person record identifiers (HHX, FMX and FPX, respectively), this program merges data from the 2005 Household, Family, Person, and Sample Child data files.

/* Merge the Household file and the Family file. */

/* Create a Household dataset with selected variables and sorted by HHX.*/
DATA HH (KEEP=HHX REGION); /* HH is a SAS dataset; the KEEP statement retains only the listed variables for processing. */
SET NHIS2005.HOUSEHLD; /*The SET statement reads data from the 2005 Household file. */
PROC SORT DATA=HH; /* Sort by HHX, the household identifier. */
BY HHX;
RUN;

/* Create a Family dataset with selected variables and sorted by HHX. */
DATA FM (KEEP=HHX FMX INCGRP RAT_CAT WTFA_FAM); /* FM is a SAS dataset; the KEEP statement retains only the listed variables for processing. */
SET NHIS2005.FAMILYXX; /*The SET statement reads data from the 2005 Family file. */
PROC SORT DATA=FM; /* Sort by HHX, the household identifier. */
BY HHX;
RUN;

DATA HHFM; /* New combined dataset called HHFM */
MERGE FM (IN=FROMFM) HH ; /* Merge the newly created FM and HH files, using an IN statement.*/
BY HHX;
IF FROMFM = 1; /* The combined dataset HHFM will contain only those records that are in the Family file; the Household file's REGION variable will be appended to these records. */
PROC SORT DATA=HHFM; /* Sort by HHX and FMX, the household and family identifiers. */

BY HHX FMX;
RUN;

In the code above, the IN statement creates a temporary SAS variable (called FROMFM) that has a value of 1 if the dataset associated with the IN statement contributed to the current observation, or a value of 0 if it did not. The subsequent statement, "IF FROMFM = 1" tells SAS to retain only those observations from the Family file (called FM). For more information on IN statements in SAS, consult Delwiche and Slaughter (1998).

/* Merge the Person file and the combined Family/Household file. */

/* Create a Person file with selected variables. */
DATA PR (KEEP=HHX FMX FPX SEX AGE_P WTFA STRATUM PSU); /* PR is a SAS dataset; the KEEP statement retains only the listed variables for processing. */
SET NHIS2005.PERSONSX; /*The SET statement reads data from the 2005 Person file. */
PROC SORT DATA=PR; /* Sort by HHX and FMX, the household and family identifiers. */
BY HHX FMX;
RUN;

DATA PRHHFM; /* Combined Person, Family, and Household dataset called PRHHFM*/
MERGE PR HHFM (DROP=WTFA_FAM); /* Merge the newly created PR file and HHFM, the combined Family/Household file, by the identifiers common to both files. At this point, users may drop the Family file weight and retain only the Person file weight for person-level analyses.*/
BY HHX FMX;
PROC SORT DATA=PRHHFM; /* Sort by HHX, FMX, and FPX, the household, family, and person identifiers. */
BY HHX FMX FPX;
RUN;

/* Merge the Sample Child file and the combined Person/Family/Household file. */

/* Create a Sample Child file with selected variables. */
DATA CH (KEEP=FPX HHX FMX CASHMEV PROBRX WTFA_SC); /* CH is a SAS dataset; the KEEP statement retains only the listed variables for processing. */
SET NHIS2005.SAMCHILD; /*The SET statement reads data from the 2005 Sample Child file. */
PROC SORT DATA=CH; /* Sort by HHX, FMX, and FPX, the household, family, and person identifiers. */
BY HHX FMX FPX;
RUN;

DATA CHPRHHFM; /* Combined Sample Child, Person, Family, and Household dataset called CHPRHHFM*/
MERGE PRHHFM CH; /* Merge CH, the newly created Sample Child file, and PRHHFM, the combined Person/Family/Household file, by the identifiers common to both files.
BY HHX FMX FPX;
RUN;

Combining Years of Data

<div style="border:1px solid black">

Important Note

Variable names may change from one year to another. Users are advised to check variable names and where names differ, make certain it is appropriate to combine years of data for a given variable.

</div>

As previously mentioned, the purpose of combining or concatenating years of data (in SAS terminology) is to increase the number of observations or respondents for the same number of variables, and thus increase the precision of estimates. It is possible to combine data from successive years of the National Health Interview Survey (NHIS) when the questions remain essentially the same over the years being combined.

Combining datasets from more than one year joins them one after the other (concatenates), as opposed to merging datasets. Analysts wishing to do both – merge data from multiple files within years and combine years of data – will need to first merge the data within each single year and then concatenate the files for the selected years of data (see the preceding section on Merging Data Files).

Weights will normally need to be adjusted when combining data years. For example, if two years of NHIS data are combined, the sum of the weights will be about twice the size of the civilian noninstitutionalized population of the United States. To achieve annualized results when two years of NHIS data are combined, one method for weight adjustment is to divide each weight by two before analyzing the data.

If data from the period 1997-2005 are combined, the combined data are treated like a single year of data with a larger sample size for the purpose of variance estimation. If data from any year before 1997 are combined with data from 1997 and beyond, variance estimation is more complicated. Refer to discussion in the first part of this document for more information about variance estimation methods when combining datasets from more than one year.

The following is an example of a SAS program that will combine data files across NHIS data years. The program is written to concatenate the data from the Person files of the 2004 NHIS and the 2005 NHIS.

/*Combine data files from 2 different years. */

DATA PER_04; /* Create SAS dataset PER_04.*/
SET NHIS2004.PERSONSX /* The SET statement reads data from an existing SAS dataset, e.g., the 2004 Person file */ **(KEEP=HHX FMX FPX AGE_P SEX WTFA STRATUM PSU);** /* The KEEP statement retains only the listed variables for processing. */
RUN;

PROC SORT DATA=PER_04; /* Sort SAS dataset PER_04. */
BY HHX FMX FPX;
RUN;

DATA PER_05; /* Create SAS dataset PER_05.*/
SET NHIS2005.PERSONSX /* The SET statement reads data from an existing SAS dataset, e.g., the 2005 Person file */ **(KEEP=HHX FMX FPX AGE_P SEX WTFA STRATUM PSU);** /* The KEEP statement retains only the listed variables for processing. */
RUN;

PROC SORT DATA=PER_05; /* Sort SAS dataset PER_05. */
BY HHX FMX FPX;
RUN;

DATA COMBO; /* New, combined SAS dataset */
SET PER_04 PER_05; /* Concatenate selected variables from 2004 and 2005 datasets. */
WTFA_2YR=WTFA/2; /*Create a new weight by dividing the existing Person file weight (WTFA) by 2, the number of Person data files combined to create the data file called COMBO.*/
RUN;

Reference

Delwiche, LD and SJ Slaughter (1998), *The Little SAS Book: A Primer* (2nd edition), SAS Institute: Cary, NC.