Variance Estimation and Other Analytic Issues in the 1997-2005 NHIS
(Adapted from Appendix III of the 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 are adjusted in several ways to yield estimates for the civilian, noninstitutionalized population of the United States. As with any variance estimation methodology, those presented here involve several simplifying assumptions about the design and weighting scheme applied to the data. This appendix provides guidelines for data users based on simplified concepts of the NHIS sample design structure so that users may compute reasonably accurate standard errors.

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 use the software package SUDAAN® (Shah et al. 1997) to produce accurate standard errors. In this appendix, examples of SUDAAN computer code are provided for illustrative purposes. 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 survey data.

✍ **CAUTION. Users are reminded that the use of standard statistical procedures for survey data analysis, which are based on the assumption that** data are generated via simple random sampling (SRS), 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 described below. Analysts who apply SRS techniques to NHIS data will produce standard errors that are, on average, too small, and are likely to produce results that are more subject to Type I error. For example, confidence intervals produced under the SRS assumption will typically be narrower than those produced taking into account the clustering and other complexities of the NHIS sample.

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 appendix, but are provided elsewhere (NCHS 1999; NCHS 2000). This appendix outlines the basic technical ideas published in these technical reports (NCHS 1999; NCHS 2000).

To achieve sampling efficiency and to keep survey operations manageable and cost-effective, 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 obtaining a simple random sample of 50,000 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 1900 Primary Sampling Units (PSUs), single counties or groups of adjacent counties (or equivalent jurisdictions) and/or metropolitan areas. These PSUs vary in population size and number of jurisdictions. The PSUs with the largest populations (e.g., the New York metropolitan area) support cost-effective sampling and are sampled with certainty; consequently, they are designated as self-representing (SR) PSUs. Resource constraints required that the remaining smaller PSUs be sampled for data collection. 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 each NSR stratum, two PSUs were selected without replacement with probability proportional to population size, and the SR PSUs were selected with certainty. For some stratum with smaller population size, only one NSR PSU was drawn from a stratum.

The U.S. Census Bureau 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 are selected. These 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 inverse probabilities 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, 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 weights are adjusted according to a quarterly poststratification by 88 age/sex/race and/or ethnicity classes based on Census control totals.

Internally, NCHS uses the design and weighting information to formulate appropriate variance estimators 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. 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 the internal NCHS data have the option of accessing internal 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 Person file public use design variables utilized for variance estimation are provided in Table 1. Users should check the Dataset Documentation for exact names and locations of these variables for each of the files.

Appendix III, Table 1. Variables Used for Variance Estimation, 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 regional 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 and later NHIS public use data files are not comparable to those of previous data years. 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.

Variance Estimation Method for Public Use Data

The method described below is applicable to all 1997 and later NHIS public use data, except the Injury Episode, Injury Verbatim, and Poison Episode files (when available).

For this method of variance estimation, the NHIS sample is treated as having 339 strata, each containing two sampled PSUs. While in reality the PSUs were not duplicated, 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 standard errors than the true variance estimation method that is applied internally by analysts at NCHS (NCHS 2000). Additionally, this public use method is applicable in many of the 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 **... 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.

✍ **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-PSU 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. The 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.

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 select records (and list-wise deleting other records) is called subsetting the data. With a subsetted dataset, which 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 which has all sample design information. (See a SUDAAN manual for more information).

Strategy 1 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 produce standard errors identical to that achieved when using a full dataset with a SUBPOPN statement (see Strategy 2, below). Note that other calculations for design effects, degrees of freedom, and standardization may need to be carried out differently. Users are responsible for verifying the correctness of their results based on subsetted data.

Strategy 2 Use the SUBPOPN statement with the method described above for the full dataset:

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

Using the full dataset with the SUBPOPN statement in this example would constrain analysis to African American women only (RACE = 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.

Combining Data Years in the National Health Interview Survey

It is sometimes possible to combine data from successive years of the National Health Interview Survey (NHIS) to increase the number of responses to questions and thus increase the precision of estimates. This is possible when the questions remain essentially the same over the years being combined.

Note that if data from the 1996 and 1997 NHIS (or any data before 1997 and data from 1997 and beyond) are combined, it is possible to obtain point estimates but not variance estimates, because the coded PSU identifications are not the same. This was done for confidentiality reasons.

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, so to achieve annualized results, each weight should be divided by two before analyzing the data.

A description of how to combine NHIS data years using SAS software follows. In SAS terminology, the process of adding observations is called concatenating data, or joining data sets one after the other, as opposed to merging data sets. The purpose of merging data is to add more variables for the same number of respondents, and this is done when data files within an NHIS data year are combined. Analysts wishing to do both - combine years and use data from multiple files for the same years - will need to both concatenate and merge data.

Below is a short explanation of the SAS **SET** command used to combine multiple years of NHIS data and an example of a program that will complete the task. The program is written to combine the data from the Person files of the 1999 NHIS and the 2000 NHIS. Note that the **SET** command does not subset a data file. That is, it is not used to sort and partition an analytic file by race, age, or any other variable that can be used to group respondents. If the research question being studied requires subsetting, please refer to the procedures and caveats discussed in the "Subsetted Data Analysis" section of this Appendix.

SAS ***SET*** Command

The **SET** statement tells the SAS system to read observations from one or more SAS data sets. An analyst will use the **SET** statement when she or he wants to concatenate or join observations from existing SAS data sets into a new data set. By default, the **SET** statement reads all of the observations from the input SAS data set. For example:

DATA **output-SAS-data-set;**
SET **input-SAS-data-set;**
RUN;

```
/*  This program concatenates the 1999 NHIS and the 2000 NHIS SAS data sets for the
Person file.  */

DATA  A;     /* SAS Data set */
SET  PERSONSX;   /* The SET statement reads data from an existing SAS data
set, e.g., the 1999 Person file. */

KEEP   HHX FMX PX RAT_CAT AGE_P WTFA STRATUM PSU IHS
OTHERGOV OTHERPUB MILITARY MEDICARE MCPART MCHMO
MEDICAID MACHMD MAPCMD MAREF PRIVATE HITYPE1
PLNMGD1 HITYPE2 PLNMGD2 HITYPE3 PLNMGD3 HITYPE4
PLNMGD4 HISCOD_I HISPAN_I INCGRP CHIP;
/* The KEEP statement retains only the listed variables for processing. */
PROC SORT DATA=A; BY HHX FMX PX;   /*  Sorting SAS Data set A */

DATA B;  /* SAS Data set */
SET PERSONSX;   /* The SET statement reads data from an existing SAS data
set, e.g., the 2000 Person file. */
KEEP  HHX FMX PX RAT_CAT AGE_P WTFA STRATUM PSU IHS
OTHERGOV OTHERPUB MILITARY MEDICARE MCPART MCHMO
MEDICAID MACHMD MAPCMD MAREF PRIVATE HITYPE1
PLNMGD1 HITYPE2 PLNMGD2 HITYPE3 PLNMGD3 HITYPE4
PLNMGD4 HISPCODR HISPANCR INCGRP CHIP;
/* The KEEP statement retains only the listed variables for processing. */
PROC SORT DATA=B; BY HHX FMX PX;  /* Sorting SAS Data set B */

DATA COMBO;    /*  New, combined SAS Data set */
SET A B ;     /*  Merging selected variables from 1999 and 2000 Data sets */
```

References

Cochran, W.G. (1977), Sampling techniques (3rd ed), 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.