VARIANCE ESTIMATION FOR THE NHIS PUBLIC USE PERSON DATA, 1995 & 1996

Created April 17, 1998; revised January 8, 2009

About this document:

This document provides basic design information about the 1995-2004 NHIS design (extended through 2005) and presents methods to compute standard errors for the 1995 and 1996 person-level data. This document also presents a method to compute standard errors for pooled 1995 and 1996 data.

Contents

VARIANCE ESTIMATION FOR THE NHIS PUBLIC USE PERSON DATA, 1995

Introduction: The data collected in the NHIS are obtained through a complex sample design involving stratification, clustering, and multistage sampling, and the final weights are subject to several adjustments. Any variance estimation methodology must involve numerous simplifying assumptions about the design and weighting. We provide some oversimplified conceptual NHIS design structures that should allow users of this Public Use Data File to compute reasonably accurate standard errors.

There are several available software packages for analyzing complex samples. The Internet 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 sample survey data. Analysts at NCHS generally use the software package SUDAAN to produce accurate standard errors. 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 analysis of complex-sample survey data.

Conceptual NHIS design for 1995-2004 (extended through 2005): The U.S. Bureau of the Census partitioned the state counties or equivalents along with metropolitan areas into a universe of about 1900 Primary Sampling Units (PSUs) (note, PSUs may be combined counties) to provide the primary sampling areas for its many national surveys. For the NHIS these universe PSUs were partitioned into geographical strata at the state level. Some of the larger universe PSUs are self-representing (SR), i.e., they are in the NHIS with certainty. The other PSUs are called non-self-representing (NSR) or non-certainty PSUs. Within each state the NSR PSUs were partitioned into strata based upon similarity of PSU characteristics. Within each NSR stratum 2 PSUs were selected using Durbin's probability proportional to size (PPS) sampling method using the population as a measure of size. (In some smaller states only 1 PSU was drawn PPS.) The SR PSUs are equivalent to strata, but historically they have been referred to as PSUs. (PPS and Durbin sampling are discussed in Chapter 9A of Cochran (1977).)

Within a sampled NSR or SR PSU the geography was partitioned into smaller geographical clusters which were used to form the universe of secondary sampling units (SSUs). These SSUs were then partitioned into density strata based upon black and Hispanic population concentration as determined by the 1990 Decennial Census. An additional stratum for new construction since the 1990 Decennial Census was also created. Within each density stratum SSUs were sampled at different rates to meet different design objectives. Within each sample SSU, all households containing black or Hispanic persons were eligible for a full NHIS interview. Only a subsample of all other households were eligible for a full NHIS interview; some were "screened out" after a brief interview.

The fundamental sampling weights were created such that under ideal sampling conditions, unbiased estimators for each level of sampling are available. In practice, however, the final sampling weights were adjusted for non-response, and ratio adjusted. Furthermore, in 1995 a federal government shutdown resulted in three lost weeks of sample which resulted in further weighting adjustments. The most important adjustment was a quarterly post-stratification to 88 age/sex/race/ethnicity independent population estimates provided by the U.S. Bureau of the Census.

For variance estimation purposes, NCHS treats the NHIS as a two-stage sample. The PSU probabilities of selection are known, and the SSUs are treated as sampled with replacement within PSU density strata. Sampling weights are adjusted by postratification. With these assumptions the SUDAAN software is used to compute variances. Much of the design information, state, density strata, and Durbin probabilities can be used to identify the smaller geographical areas. NCHS forbids the disclosure of information which may compromise the confidentiality promised to survey respondents, so some design information is not provided with the Public Use Data. While all design information is not available to the public, variance estimation methods exist which provide similar results to the NCHS internally used methodology, as described below.

Design Information Available on the 1995 NHIS Public Use Data Files

The following variables are used to produce codes for variance estimation. Field locations below are from the 1995 PERSON level data file, but may be different on other data files; the user should check the file documentation.

Variable Name	Location	Field Label
STRAT_V	337-340	'STRATA FOR VARIANCE ESTIMATION'
PSU_V	341	'PSU FOR VARIANCE ESTIMATION'
SUB_V	342-343	'SUBSTRATUM FOR VARIANCEESTIMATION'
SSU	344-350	'SECONDARY SAMPLING UNIT'
PANEL	352	'PANEL 4'
TYPE_PSU	351	'TYPE OF PSU'
WTF	219-227	'FINAL BASIC WEIGHT'

Two methods of variance estimation are now provided.

Method 1 - 187 Strata containing 2 PSUs per stratum sampled with replacement

Here, the NHIS universe has been partitioned into 187 strata. Most of the original NHIS strata and PSUs retain their original sampling structure with two PSUs being sampled per stratum, but a few strata have been collapsed, and in the largest self-representing strata, two pseudo-PSUs have been created. All PSUs are treated as sampled with replacement within their respective strata. This method will provide somewhat conservative standard errors, and the standard error estimator itself has less stability than the standard error estimator described by Method 2 below. Method 1 should be applicable to many complex survey sample design computer programs which require exactly 2 sampled PSUs per stratum. This method is robust when analyzing subsetted data (See the section "Subsetted Data Analyses" below).

Coding required, (SAS code provided):

STRATUM = STRAT_V ;

PSU = PANEL ;

IF (PSU_V = 5) THEN PSU = INT((PANEL + 1) /2) ;

IF (PSU_V = 8) THEN STRATUM = 553 ;

IF ((TYPE_PSU = 1) AND (PSU_V IN (2,4))) THEN STRATUM = (STRAT_V -1) ;

IF ((STRAT_V = 921) AND (PSU_V = 3)) THEN STRATUM = 901 ;

As a check the user should observe 374 PSUs when using the full data file.

For the above simplification of the NHIS sample design structure, the following SUDAAN design statements may be used. (Note, the input file must first be sorted by STRATUM and PSU variables.)

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

See the Section "Worked SUDAAN Examples" below for further discussion.

Corresponding statements for other software packages are as follows:

Stata svy:

SVYSET **[PWEIGHT=WTF]**,*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=***WTF**
*/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 **WTF**;
VAR <name of variable to be analyzed>;
RUN;

PROC SURVEYFREQ;
STRATA **STRATUM**;
CLUSTER **PSU**;
WEIGHT **WTF**;
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=~**wtf**,
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 **WTF**

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.

Method 2 - Multiple PSUs per Stratum design sampled with replacement

This method provides for more statistically efficient variance estimation than Method 1, since it makes better use of the sampling design information. Its application is limited to software that can handle multiple PSUs per stratum, e.g., SUDAAN. For this method the original certainty PSUs are partitioned by aggregations of the original race-ethnic density strata used in sampling. The first randomly sampled unit is actually the SSU variable which is now treated as the PSU variable. (Note, a certainty PSU unit contributes nothing to the variance at the PSU sampling level.) Non-certainty-strata PSUs are treated as being sampled with replacement within their respective strata. Except for a few special cases, the non-certainty PSUs have exactly the same structure in both Methods 1 and 2.

Coding required, (SAS code provided):

```
IF TYPE_PSU = 1 THEN DO ;    /* certainty strata PSUs */
        STRATUM2 = STRAT_V*1000 + SUB_V ;
        PSU2 = SSU ;
END;

ELSE DO ;                /* non-certainty PSU */ ;
        STRATUM2 = STRAT_V ;
        PSU2 = PSU_V ;
END;
```

As a check, the user should observe the following counts:

Certainty Strata PSUs	4079
Non-certainty Strata PSUs	259
Total PSUs	4338

For the Method 2 design structure, the following SUDAAN design statements may be used. (Note, the input file must first be sorted by STRATUM2 and PSU2 variables.)

PROC ... **DESIGN = WR ;**
NEST **STRATUM2 PSU2 ;**
WEIGHT **WTF ;**

See the Section "Worked SUDAAN Examples" for further discussion.

CAUTION. Method 2 should only be used on the full sample person data file. Using this method with subsetted data may lead to incorrectly computed standard errors. (See the section "Subsetted Data Analyses" below.) If using a subsetted data set, the user should check the degree of agreement of the certainty and non-certainty counts with the values presented above.

CAUTION

A typically used rule-of-thumb for degrees of freedom to associate with a standard error is the quantity (number of PSUs - number of strata). This rule assumes that the PSUs are somewhat comparable in size. For Method 2 this rule may be grossly inaccurate since the concept of PSU is quite different for certainty and non-certainty strata. Certainty strata PSUs of Method 2 have small weighted values relative to those of non-certainty PSUs. The rule-of-thumb degrees of freedom for Method 1 is 187, and Method 2 should have a "true" degrees of freedom exceeding that of Method 1. However, for practical purposes, any degrees of freedom exceeding 120 can be treated as infinite, i.e., one uses a normal Z-statistic instead of a t-statistic for testing. Note that a one-tailed critical $t_{0.025}$ at 120 degrees of freedom is 1.98 while at an infinite degrees of freedom (i.e., a z-value) is 1.96. If a variable of interest covers most of the NHIS PSUs, the limiting value would probably be adequate for analysis. The user should consult a mathematical statistician for discussion of degrees of freedom.

SUBSETTED DATA ANALYSES

Frequently, studies of NHIS variables are restricted to select subpopulations, e.g., persons aged 65 and older. To save on storage the user may delete all records outside of the domain of interest. This procedure of keeping only select records is called subsetting the data. With a subsetted data set one can produce correct point estimates, e.g., the subpopulation means, but standard errors may be computed incorrectly when using a compromised design structure. For example, if a stratum of Method 2 contains 10 PSUs and 5 are lost because of subsetting, a SUDAAN run on the subsetted data will use an incorrect formula to compute stratum contributions to the variance. If the full data are run, SUDAAN correctly handles the 5 empty PSUs. Note, that SUDAAN has a SUBPOPN option that allows the targeting of a subpopulation from a full design data file. (See a SUDAAN manual for details.) **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**.

<u>Subsetting methods with SUDAAN</u>

Strategy 1 (recommended): Use Method 1 above for the full data file, and the SUBPOPN statement to identify the subpopulation of interest. For example, if the subpopulation of interest is persons aged 65 and older:

SUBPOPN* AGE GE 65 *;

Strategy 2 (not recommended, except when Strategy 1 is infeasible): Use Method 1 above

with the MISSUNIT option on the NEST statement:

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

In a WR design with exactly 2 PSUs per stratum, when some PSUs are removed from the data file then the SUDAAN MISSUNIT option "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 design effects, degrees of freedom, standardization, etc. may need to be carried out differently. The user is 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(**AGE>=65**)*: MEAN* <name of variable to be analyzed>

SPSS csdescriptives or cstabulate:

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

```
DO IF (AGE GE 65).
  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 AGE >= 65        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 age>=65 without using equal signs
subgrp <- *subset*(*nhissvy*,(age>64))
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 AGE INTO AGECAT
CLASS AGECAT (LOW-64/65-HIGH)

The second category of AGECAT 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* AGECAT(2)

Note that the specification of AGECAT(2) refers to the second category of AGECAT, which is defined as all values of AGE equal to 65 and all higher values of age that occur in the data.

<u>Other notes on Subsetting data</u>:

The condition, doctor visit, and hospital data files are actually subsetted files. To use with SUDAAN or other complex design variance estimation software package properly, the information should be linked back to the appropriate person on the person file. Consult with a statistician for appropriate usage of SUDAAN and similar software.

WORKED SUDAAN EXAMPLES

In the following runs the variables used are:

LDR = proportion of persons without a doctor visit in the last 2 years

TDV_R = mean number of annual doctor visits (based upon 2 week recall)

HLT_FP = proportion of persons with self-reported fair or poor health status (omitting missing)

AGE2: 1 = aged less than 18
2 = aged 18 to 44
3 = aged 45 to 64
4 = aged 65 and older

The following SUDAAN code was executed for both Method 1 and Method 2:

Caution The output presented below was based upon a preliminary 1995 NHIS Public Use data file. The final Public Use data file may produce slightly different SUDAAN output.

```
PROC DESCRIPT DATA = HIS.infile FILETYPE=SAS DESIGN = WR ;
NEST <stratum variable> <PSU variable> ;
WEIGHT WTF ;
VAR LDR TDV_R HLT_FP ;
SUBGROUP   SEX    AGE2 ;
LEVELS     2      4 ;
TABLES SEX AGE2 ;
PRINT NSUM WSUM MEAN SEMEAN
      / WSUMFMT=F10.0 MEANFMT=F8.5 SEMEANFMT=F8.5 ;
RUN;
```

Method 1: partial output:

```
                              S U D A A N
            Software for the Statistical Analysis of Correlated Data
           Copyright      Research Triangle Institute      April 1996
                              Release 7.00


Number of observations read    : 102467    Weighted count :261889548
Number of observations skipped :      0
(WEIGHT variable nonpositive)
Denominator degrees of freedom :    187
```

Research Triangle Institute
The DESCRIPT Procedure

by: Variable, SEX.

Variable		SEX Total	1	2
LDR	Sample Size	102467	48809	53658
	Weighted Size	261889549	127570237	134319312
	Mean	0.13797	0.18013	0.09793
	SE Mean	0.00178	0.00250	0.00178
TDV_R	Sample Size	102467	48809	53658
	Weighted Size	261889549	127570237	134319312
	Mean	5.90759	4.90385	6.86089
	SE Mean	0.09060	0.10039	0.12407
HLT_FP	Sample Size	101277	48266	53011
	Weighted Size	258963568	126221708	132741859
	Mean	0.10126	0.09124	0.11079
	SE Mean	0.00157	0.00188	0.00176

by: Variable, AGE2.

Variable		AGE2 Total	1	2	3	4
LDR	Sample Size	102467	29711	40801	20000	11955
	Weighted Size	261889549	70670755	108040689	51713265	31464840
	Mean	0.13797	0.08894	0.18489	0.14461	0.07606
	SE Mean	0.00178	0.00269	0.00268	0.00293	0.00251
TDV_R	Sample Size	102467	29711	40801	20000	11955
	Weighted Size	261889549	70670755	108040689	51713265	31464840
	Mean	5.90759	4.29682	4.88589	7.08504	11.09843
	SE Mean	0.09060	0.09797	0.12432	0.17859	0.30642
HLT_FP	Sample Size	101277	29183	40423	19834	11837
	Weighted Size	258963568	69438212	107054300	51315866	31155190
	Mean	0.10126	0.02552	0.06610	0.16651	0.28344
	SE Mean	0.00157	0.00129	0.00168	0.00356	0.00519

Method 2: Partial Output:

```
                              S U D A A N
           Software for the Statistical Analysis of Correlated Data
          Copyright      Research Triangle Institute      April 1996
                              Release 7.00

Number of observations read    : 102467    Weighted count :261889548
Number of observations skipped :      0
(WEIGHT variable nonpositive)
Denominator degrees of freedom :   4030


                           Research Triangle Institute
                             The DESCRIPT Procedure


by: Variable, SEX.
```

Variable		SEX Total	1	2
LDR	Sample Size	102467	48809	53658
	Weighted Size	261889549	127570237	134319312
	Mean	0.13797	0.18013	0.09793
	SE Mean	0.00174	0.00231	0.00184
TDV_R	Sample Size	102467	48809	53658
	Weighted Size	261889549	127570237	134319312
	Mean	5.90759	4.90385	6.86089
	SE Mean	0.07704	0.08503	0.11403
HLT_FP	Sample Size	101277	48266	53011
	Weighted Size	258963568	126221708	132741859
	Mean	0.10126	0.09124	0.11079
	SE Mean	0.00152	0.00174	0.00182

by: Variable, AGE2.

Variable		AGE2 Total	1	2	3	4
LDR	Sample Size	102467	29711	40801	20000	11955
	Weighted Size	261889549	70670755	108040689	51713265	31464840
	Mean	0.13797	0.08894	0.18489	0.14461	0.07606
	SE Mean	0.00174	0.00271	0.00254	0.00303	0.00269
TDV_R	Sample Size	102467	29711	40801	20000	11955
	Weighted Size	261889549	70670755	108040689	51713265	31464840
	Mean	5.90759	4.29682	4.88589	7.08504	11.09843
	SE Mean	0.07704	0.09116	0.11805	0.16109	0.28387
HLT_FP	Sample Size	101277	29183	40423	19834	11837
	Weighted Size	258963568	69438212	107054300	51315866	31155190
	Mean	0.10126	0.02552	0.06610	0.16651	0.28344
	SE Mean	0.00152	0.00118	0.00157	0.00351	0.00501

Best NHIS design using Durbin probabilities (not available to the public) and weights adjusted by post-stratification:

Research Triangle Institute
The DESCRIPT Procedure

Post-stratified estimates
by: Variable, SEX.

Variable		SEX Total	1	2
LDR	Sample Size	102467	48809	53658
	Weighted Size	261889549	127570237	134319312
	Mean	0.13784	0.17991	0.09789
	SE Mean	0.00170	0.00221	0.00182
TDV_R	Sample Size	102467	48809	53658
	Weighted Size	261889549	127570237	134319312
	Mean	5.90468	4.89733	6.86141
	SE Mean	0.07511	0.08320	0.11217
HLT_FP	Sample Size	101277	48266	53011
	Weighted Size	258974266	126232939	132741328
	Mean	0.10127	0.09125	0.11080
	SE Mean	0.00137	0.00159	0.00165

Post-stratified estimates
by: Variable, AGE2.

Variable		AGE2 Total	1	2	3	4
LDR	Sample Size	102467	29711	40801	20000	11955
	Weighted Size	261889549	70670755	108040689	51713265	31464840
	Mean	0.13784	0.08845	0.18484	0.14484	0.07587
	SE Mean	0.00170	0.00258	0.00248	0.00298	0.00268
TDV_R	Sample Size	102467	29711	40801	20000	11955
	Weighted Size	261889549	70670755	108040689	51713265	31464840
	Mean	5.90468	4.29787	4.87876	7.08472	11.09687
	SE Mean	0.07511	0.09066	0.11858	0.16180	0.27613
HLT_FP	Sample Size	101277	29183	40423	19834	11837
	Weighted Size	258974266	69441900	107059972	51315313	31157082
	Mean	0.10127	0.02555	0.06624	0.16633	0.28322
	SE Mean	0.00137	0.00116	0.00153	0.00342	0.00487

Remark on Examples

A comparison of the three SUDAAN examples shows that Method 2 performs quite well when compared to the "best" internal NCHS variance design for the NHIS. Based on limited preliminary evidence, it appears that for means, Method 2 typically provides standard errors in close agreement with, while slightly larger than, the standard errors produced by the NCHS "best" method. Method 1 tends to provide slightly larger standard errors than Method 2 does, although the sample output does include examples where the Method 1 standard error is smaller than the Method 2 standard error.

Reference:

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

Notes for the 1995 Year 2000 Objectives Public Use File

The variance estimation methods of this document may be applied to the 1995 Year 2000 Objectives Public Use File. The following changes must be made:

The design information variables are all in the same file locations with the exception of "WTF". Substitute:

WTF	207-212	'FINAL BASIC WEIGHT'

The PSU check for **method 2** should now read:

As a check, the user should observe the following counts:

Certainty Strata PSUs	3804
Non-certainty Strata PSUs	259
Total PSUs	4063

VARIANCE ESTIMATION FOR THE NHIS PUBLIC USE PERSON DATA, 1996

In 1996 the NHIS began the transition from a paper-and-pencil questionnaire to a computer-assisted interview process. This transition resulted in roughly 5/8 of the full 1996 sample being available for processing and public release. In 1997 the full NHIS sample was administered using computer-assisted interviewing. NCHS created a different variance estimation structure for 1996 than for 1995, which is described below.

Design Information Available on the 1996 NHIS Public Use Data Files

The following variables are used to produce code for variance estimation. Field locations below are from the 1996 PERSON level data file, but may be different on other data files; the user should check the file documentation.

Variable Name	Location	Field Label
STRAT96*	**354-357**	**'COLLAPSED VARIANCE STRATUM'**
PSU96*	**358**	**'VARIANCE PSU'**
SUB_V	342-343	'SUBSTRATUM FOR VARIANCE ESTIMATION'
SSU	344-350	'SECONDARY SAMPLING UNIT'
PANEL	352	'PANEL 4'
NSR96*	**353**	**'NSR STATUS VARIABLE'**
WTF	219-227	'FINAL BASIC WEIGHT'

(***** indicates modified design variables provided with the 1996 data files)

Two methods of variance estimation are now provided.

Method 1.96 - 98 Strata containing 3 PSUs per stratum sampled with replacement

Here, the NHIS universe has been partitioned into 98 collapsed strata with 3 PSUs per stratum. All PSUs are treated as sampled with replacement within their respective strata. This method will provide somewhat conservative standard errors, and this standard error estimator itself has less stability than the standard error estimator described by Method 2.96 below.

Coding required, (SAS code provided):

```
STRATUM96 = INT( STRAT96 / 10 ) * 10 ;

PSU96 = PANEL ;
```

Note**, INT ( )** is the Integer-value SAS function which truncates the decimal part of the input value, e.g., INT(2.3) = 2

As a check the user should observe 98*3 = 294 PSUs when using the full data file.

For the above simplification of the NHIS sample design structure, the following SUDAAN design statements may be used. (Note, the input file must first be sorted by the STRATUM96 and PSU96 variables.)

***PROC* ... DESIGN = WR;**
***NEST* STRATUM96 PSU96 ;**
***WEIGHT* WTF ;**

Corresponding statements for other software packages: refer to example code for 1995 Method 1, substitute "STRATUM96" and "PSU96" for "STRATUM" and "PSU", respectively. Similarly, for subsetted data analyses, refer to example code for 1995 Method 1, substituting "STRATUM96" and "PSU96" for "STRATUM" and "PSU", respectively.

Method 2.96 - Multiple PSUs per Stratum design sampled with replacement

This method provides for more statistically efficient variance estimation than Method 1.96, since it makes better use of the sampling design information. Its application is limited to software that can handle multiple PSUs per stratum, e.g., SUDAAN. For this method the original certainty PSUs are partitioned by aggregations of the original race-ethnic density strata used in sampling. The first randomly sampled unit is actually the SSU variable which is now treated as the PSU variable. (Note, a certainty PSU unit contributes nothing to the variance at the PSU sampling level.) Non-certainty-strata PSUs are treated as being sampled with replacement within their respective strata.

Coding required, (SAS code provided):

```
IF NSR96 = 1 THEN DO ;    /* 1996 certainty strata PSUs    */
        STRATUM96_2 = STRAT96*100 + SUB_V ;
        PSU96_2 = SSU ;
END;

ELSE DO ;                 /* 1996 non-certainty PSU */ ;
        STRATUM96_2 = STRAT96 ;
        PSU96_2 = PSU96 ;
END;
```

As a check, the user should observe the following counts:

Certainty Strata PSUs	1736
Non-certainty Strata PSUs	240
Total PSUs	1976

For the Method 2.96 design structure, the following SUDAAN design statements may be used. (Note, the input file must first be sorted by STRATUM96_2 and PSU96_2 variables.)

PROC ... **DESIGN = WR ;**
NEST **STRATUM96_2 PSU96_2 ;**
WEIGHT **WTF ;**

CAUTION. Both Method 1.96 and Method 2.96 should only be used on a full sample person data file. Using this method with subsetted data may lead to incorrectly computed standard errors. (See the section "Subsetted Data Analyses" in the 1995 section).

CAUTION
A typically used rule-of-thumb for degrees of freedom to associate with a standard error is the quantity (number of PSUs - number of strata). This rule assumes that the PSUs are somewhat comparable in size. For Method 2.96 this rule may be grossly inaccurate since the concept of PSU is quite different for certainty and non-certainty strata. Certainty strata PSUs of Method 2.96 have small weighted values relative to those of non-certainty PSUs. The rule-of-thumb degrees of freedom for Method 1.96 is 196, and Method 2.96 should have a "true" degrees of freedom exceeding that of Method 1.96. However, for practical purposes, any degrees of freedom exceeding 120 can be treated as infinite, i.e., one uses a normal Z-statistic instead of a t-statistic for testing. Note, that a one-tailed critical $t_{0.025}$ at 120 degrees of freedom is 1.98 while at an infinite degrees of freedom (i.e., a z-value) is 1.96. If a variable of interest covers most of the NHIS PSUs, the limiting value would probably be adequate for analysis. The user should consult a mathematical statistician for discussion of degrees of freedom.

The observant reader may notice that the 1996 method 1.96 has a larger "rule of thumb" degrees of freedom than the corresponding 1995 method 1. The 1996 variance estimation design consists of collapsed strata that may introduce a much larger stratum-collapse bias than occurred in 1995, and furthermore, the PSUs within each 1996 collapsed stratum have greater PSU weight diversity than in 1995 which may reduce stability.

The section on **SUBSETTED DATA ANALYSES** in the 1995 section should be read considering the changes provided in this 1996 section.

VARIANCE ESTIMATION FOR POOLED 1995 AND 1996 NHIS DATA

The 1995 and 1996 structures described above are not compatible with each other, and thus are not suitable for an analysis involving pooled 1995 and 1996 NHIS data. The following structure is compatible across 1995 and 1996. A limited empirical investigation suggested that this structure gives slightly higher standard error estimates than the 1995 Method 1 structure, and standard error estimates comparable to the 1996 Method 1.96 structure.

Design Information Available on both the 1995 and 1996 NHIS Public Use Data Files

Field locations below are from the 1995 and 1996 PERSON level data files, but may be different on other data files; the user should check the file documentation.

Variable Name	Location	Field Label
STRAT_V	337-340	'STRATA FOR VARIANCE ESTIMATION'
PANEL	352	'PANEL 4'

Coding required, (SAS code provided):

```
STRAT9596 = INT( STRAT_V / 10 );

PSU9596 = PANEL ;

IF STRAT9596=100 THEN STRAT9596=99;
```

As a check the user should observe 99 strata and 393 PSUs when using the full 1995 data file, and 99 strata and 295 PSUs when using the full 1996 data file.

For this structure, the following SUDAAN design statements may be used. (Note, the input file must first be sorted by STRAT9596 and PSU9596 variables.)

```
PROC ...      DESIGN = WR;
NEST STRAT9596 PSU9596 ;
WEIGHT WTF ;
```

Note that survey year is not part of the NEST variable specification, because the two years of survey data are not independent.

Corresponding statements for other software packages: refer to example code for 1995 Method 1, substitute "STRAT9596" and "PSU9596" for "STRATUM" and "PSU", respectively. Similarly, for subsetted data analyses, refer to example code for 1995 Method 1, substituting "STRAT9596" and "PSU9596" for "STRATUM" and "PSU", respectively.