VARIANCE ESTIMATION FOR THE 1985-94 NHIS PUBLIC USE PERSON DATA

Introduction:

This document presents two methods for computing standard errors for the 1985-94 NHIS person-level data. Method 1 applies to the entire 1985-94 period. It may be used for subsetted data analyses, and it is suitable for analyses of pooled data in the 1985-94 period. Method 2 applies only to the 1987-94 period and is not recommended for subsetted data analyses.

Design Information Available on the 1985-94 NHIS Public Use Data Files

The following variables are used to produce codes for variance estimation. Field locations below are the 1985-94 PERSON level data file, but may be different on other data files; the user should check the file documentation.

Variable Name	Location	Field Label
STRATUM	179-181	'FULL SAMPLE STRATUM IDENTIFIER' (Not available on 1985-86 files)
CSTRATUM	187-188	'PSEUDO PSU CODES', first two columns Values for 1985-94: 1, 2, ..., 62
CPSU	189	'PSEUDO PSU CODES', last column Values for 1987-94: 1, 2, 3, 4 Values for 1985: 1, 2, 3 Values for 1986: 2, 3
SUB	178	'TYPE OF SUBSTRATUM' Values for 1987-94: 0, 1, 2 (Not available on 1985-86 files)
SSU	5-12	concatenation of 'PROCESSING QUARTER', 'RANDOM RECODE OF PSU NUMBER', 'WEEK-CENSUS CODE' and 'SEGMENT NUMBER'
TYPE_PSU	185	'TYPE OF PSU' Values: Self-representing PSU: 1, 4 Non-self-representing PSU: 3, 6
WTF	207-212	'FINAL BASIC WEIGHT'

Method 1 - Single Stage PSUs Sampled With-Replacement within Strata Design for the 1985-94 NHIS

This method is statistically less efficient than the method described below but is more flexible.

This method requires no recoding of design variables, may be applicable to many complex survey sample design computer programs, and covers the 1985-94 NHIS survey years. Using the variables CSTRATUM, CPSU, and WTF, the PSU unit CPSU is treated as being sampled with replacement within stratum unit CSTRATUM. The data file needs to be sorted only by CSTRATUM and PSU prior to using SUDAAN.

For the above simplification of the NHIS sample design structure, use the following SUDAAN design statements:

```
PROC ...        DESIGN = WR;
NEST CSTRATUM CPSU;
WEIGHT WTF;
```

Corresponding statements for other software packages are as follows:

Stata svy:

```
SVYSET [PWEIGHT=WTF],STRATA(CSTRATUM)PSU(CPSU)
SVY: MEAN <name of variable to be analyzed>
```

SPSS csdescriptives:

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=CSTRATUM CLUSTER=CPSU
/ESTIMATOR TYPE=WR.
```

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

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

SAS proc surveymeans :

```
PROC SURVEYMEANS;
STRATA CSTRATUM;
CLUSTER CPSU;
WEIGHT WTF;
VAR <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=~cpsu, strata=~cstratum,
                     nest = TRUE,
                     weights=~wtf,
                     data=< existing data frame name>)
svymean(~<name of variable to be analyzed>,design=nhissvy)
```

VPLX:

In the CREATE step, include the following statements:

```
STRATUM    CSTRATUM
CLUSTER    CPSU
WEIGHT     WTF
```

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

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

Method 2 - Multi-stage stratified sampling design for the 1987-94 NHIS.

This design provides for more statistically efficient variance estimation than Method 1, since it makes fewer simplifications of the NHIS sample design structure but is only applicable to SUDAAN. This method also requires recodes of the design variables and is only applicable to survey years 1987-94 NHIS data.

Prior to use of this method the following recoding must be done on the NHIS file. This example is in SAS but other programming languages may be used.

```
If (TYPE_PSU = 1 or TYPE_PSU = 4) then do;
        PSU = 1;
        POPPSU = 0;
END;
If (TYPE_PSU = 3 or TYPE_PSU = 6) then do;
        PSU = CPSU;
        POPPSU = -1;
END;
```

On the record for each person, this recode creates two new variables PSU and POPPSU for use by SUDAAN's NEST and TOTCNT statements. For more information on the purpose of these statements refer to SUDAAN documentation. With these additional variables the following describes SUDAAN code for NHIS data-sets assuming a multi-stage stratified sampling design.

Before running SUDAAN against the data file, however, sort the input file by the NEST variables (STRATUM, PSU, SUB, and SSU).

For SUDAAN describe the NHIS sampling plan as follows:

***PROC* ...   DESIGN = WOR;**
***NEST* STRATUM PSU SUB SSU /MISSUNIT;**
***TOTCNT* POPPSU _ZERO_ _MINUS1_ _ZERO_;**

Caution. This method assumes that ALL records on the public-use data file are being used. This method is not recommended for use with subsetted data.

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 because some of the sample design information is unavailable to the variance estimation software. **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**.

Subsetting methods with SUDAAN

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* CSTRATUM CPSU/ 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:

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, e.g.:

```
CSDESCRIPTIVES
/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:

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;
```

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.