January 8, 2009

VARIANCE ESTIMATION FOR THE 1963-72 NHIS PUBLIC USE PERSON DATA

Introduction:

This document presents a method for computing standard errors for the 1963-72 NHIS person-level data. It may be used for subsetted data analyses, and it is suitable for analyses of pooled data in the 1963-72 period.

Variance Estimation Method: Single Stage PSUs Sampled With Replacement within Strata Design for 1963-72 NHIS.

NCHS has created a set of three-digit codes for variance estimation. The codes for 1963-67 and 1970-72 take values 001, 002, ... 286. The codes for 1968 and 1969 are almost the same; the only difference is that codes 109 and 110 are not present. Each code value denotes a distinct pseudo-PSU for variance estimation, and each pair of code values (e.g., (001,002), (003,004), ... (285,286)) should be grouped together into a pseudo-stratum for variance estimation. That is, for 1963-67 and 1970-72 NHIS public-use file variance estimation, there are 286 pseudo-PSUs grouped into 143 pseudo-strata, two pseudo-PSUs per pseudo-stratum. For 1968 and 1969 NHIS public use file variance estimation, there are 284 pseudo-PSUs grouped into 142 pseudo-strata.

Pseudo-stratum numbers 001, 002, ... 143 have been included in this file for the convenience of the user. Pseudo-stratum number 055 is not present for 1968 and 1969.

NCHS has created a variance estimation file for the 1968-72 calendar year NHIS public-use files with the following layout:

COLUMNS	FIELD
1-2	Year (e.g., 68, 69, 70, 71, 72)
3-5	Random Recode of PSU
6-7	Week
8-9	Segment
10-12	Pseudo-stratum (SUDOSTR)
13-15	Pseudo-PSU (SUDOPSU)

The file is in ASCII format, and it is sorted by Year, Random Recode of PSU, Week, and Segment.

A similar file is available for the 1963-68 fiscal year NHIS public use files with the same record layout.

To assign pseudo-stratum/pseudo-PSU codes to an NHIS file for a given calendar year, sort the NHIS file by Random Recode of PSU, Week, and Segment, and match the 1968-72 NHIS variance estimation file records for that calendar year to the NHIS file. There should be no mismatches. Follow the same process for the fiscal year 1963-68 files.

The NHIS file then should be sorted by pseudo-stratum (SUDOSTR) and pseudo-PSU (SUDOPSU) prior to invoking SUDAAN.

Use the following SUDAAN design statements:

PROC *<DESCRIPT, CROSSTAB, ...>***...** **DESIGN = WR;**
NEST **SUDOSTR SUDOPSU**;
WEIGHT **< weight variable name >;**

Corresponding statements for other software packages are as follows:

Stata svy:

SVYSET **[PWEIGHT=< weight variable name >]**,*STRATA***(SUDOSTR)***PSU***(SUDOPSU)**
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=< **weight variable name** >
*/DESIGN STRATA=***SUDOSTR** *CLUSTER=***SUDOPSU**
*/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 **SUDOSTR**;
CLUSTER **SUDOPSU**;
WEIGHT **< weight variable name >**;
VAR <name of variable to be analyzed>;
RUN;

PROC SURVEYFREQ;
STRATA **SUDOSTR**;
CLUSTER **SUDOPSU**;
WEIGHT **< weight variable name >**;
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*=~**sudopsu**, *strata*=~**sudostr**,
nest = TRUE,
weights=~**< weight variable name >**,
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 **SUDOSTR**
CLUSTER **SUDOPSU**
WEIGHT **< weight variable name >**

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.

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 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 the MISSUNIT option on the NEST statement:

***NEST* SUDOSTR SUDOPSU/ 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.