VARIANCE ESTIMATION FOR PERSON DATA USING THE
NHIS PUBLIC USE PERSON DATA TAPE, 1995

January 21,1998

(See Notes at end of this document for application to the Year 2000 Objectives data file.)

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 Set to compute reasonably accurate standard errors.

There are several available software packages for analyzing complex samples. A comparison is beyond the scope of this document, but an Internet web page *Summary of Survey Analysis Software* currently located at **http://www.fas.harvard.edu/~stats/survey-soft/survey-soft.html** provides references and discussion. At NCHS the software package SUDAAN® has been used to produce standard errors. In this document SAS® and SUDAAN® computer code is provided, but without guarantees of any kind. The computer code and methods are subject to change without notification to the user. The entire risk as to the results and performance is assumed by the user. NCHS recommends that any analysis of NHIS data be done under the supervision of a statistician who understands the implications of complex-sample design surveys.

Conceptual NHIS design for 1995 The U.S. Bureau of the Census partitions 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 are 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 are partitioned into strata based upon similarity of PSU characteristics. Within each NSR stratum 2 PSUs are 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 is 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 is partitioned into smaller geographical clusters which are used to form the universe of secondary sampling units (SSUs). These SSUs are then partitioned into density strata based upon black and Hispanic population concentration as determined by the 1990 Decennial Census. An additional strata for new construction since the last Decennial Census is also created. Within each density stratum SSUs are sampled at different rates to meet different design objectives. Within each sample SSU, all households containing black or Hispanic persons are sampled, while all other households are subsampled. Supplemental NHIS surveys may require additional sampling at SSU, household, or family levels.

The fundamental sampling weights are created such that under ideal sampling conditions, unbiased estimators for each level of sampling are available. In practice, however, the final sampling weights are adjusted for non-response, and ratio adjusted. Furthermore, in 1995 a government shutdown resulted in three lost weeks of sample which resulted in further weighting adjustments. The most important adjustment is a quarterly post-stratification to 90 age/sex/race/ethnicity Census control totals.

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. Two methods are described below.

Design Information Available on the NHIS Public Use Databases.

The following variables are used to produce code for variance estimation. Field locations below are from the PERSON level database, but may change on other databases; the user should check the file documentation.

Variable Name	Tape Location	Field Label
STRAT_V	337-340	'STRATA FOR VARIANCE ESTIMATION'
PSU_V	341	'PSU FOR VARIANCE ESTIMATION'
SUB_V	342-343	'SUBSTRATUM FOR VARIANCE ESTIMATION'
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 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 Analysis" 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 database.

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 ...   DESIGN = WR;
NEST       STRATUM PSU ;
WEIGHT     WTF ;
```

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

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    */

    STRATUM = STRAT_V*1000 + SUB_V ;
          PSU   = SSU  ;
          END ;

 ELSE DO ;                  /* non-certainty PSU */  ;

    STRATUM = STRAT_V ;
          PSU  =   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 STRATUM and PSU variables.)

```
PROC ...   DESIGN = WR;
NEST       STRATUM PSU ;
WEIGHT     WTF ;
```

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

☹**CAUTION**. Method 2 should only be used on a full sample person data base. Using this method with subsetted data may lead to incorrectly computed standard errors. (See the section "Subsetted Data Analysis" 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 subdomains, 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 subdomain 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 subdomain from a full design data base. (See the SUDAAN manual for details).

Subsetting methods with SUDAAN

Strategy 1. Use Method 1 above with the MISSUNIT option on the NEST statement -

NEST **STRATUM PSU/ MISSUNIT ;**

If a WR design has exactly 2 PSUs per stratum and some PSUs are removed from the database then the SUDAAN MISSUNIT option performs a fix-up which produces a standard error identical to that achieved when using a full data set and SUBPOPN statement. Note, other output like design effects, degrees of freedom, standardization may be computed differently. The user is responsible for checking that subsetted input leads to correct results.

Strategy 2. Use Method 1 or 2 above on a "fixed-up" subsetted data set. Basically, one needs to add some dummy records containing full design information to the subsetted data set. To do this follow these instructions:

1. Create a 2-variable file containing STRATUM and PSU for each record of the full person file (100,000+ records)

2. Sort this file by STRATUM and PSU within STRATUM.

3. Keep only 1 record for each PSU
   add WTF = 10^{-10} as a very small weight
   add variable DUMMY = 0 to designate dummy record

   A file, called DESIGN containing 4 variables with
   374 records (Method 1 used) or with
   4338 records (Method 2 used) is created

4. Append DESIGN to the original subsetted database, called DATASET, to form a new set, called DATANEW.

Define DUMMY = 1 on the DATASET component.

On the DESIGN component records define all variables other than STRATUM, PSU, WTF, DUMMY as missing "." .

5. Sort DATANEW by STRATUM PSU

6. In SUDAAN use a "*SUBPOPN* DUMMY = 1;" line to direct SUDAAN to restrict estimation to the subdomain of interest.

With the above fix-up SUDAAN will correctly handle empty PSUs when computing the standard errors. SUDAAN output that needs the entire full sample database for correct computation, e.g, design effects, may or may not be appropriate. See the SUDAAN manual for computational forms or consult with a mathematical statistician for correct interpretation.

Other notes on Subsetting data:

If a subsetted database under Method 2 has only a few missing PSUs , the subsetted database can probably be run with SUDAAN without being fixed up. For example, a subsetting by SEX will most likely result in all PSUs still being in sample, but black males aged 65 and older would result in the loss of many PSUs. The impact of running SUDAAN on uncorrected subsetted data varies. Frequently, subsetted runs produce results consistent with those run on a full data set, but sometimes they do not.

Subsetting by aggregates of Strata does not need a fix-up.

The condition, doctor visit, and hospital record databases are actually subsetted files. To use with SUDAAN properly, the information should be linked back to the appropriate person on the person file. Some statistics, based upon aggregation of records, may be computed directly from this file along with the fix-up. Consult with a statistician for appropriate SUDAAN usage.

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 is based upon a preliminary NHIS Public Use database. Your Public Use database may produce slightly different SUDAAN output.

```
PROC DESCRIPT  DATA = HIS.infile  FILETYPE=SAS DESIGN = WR;

NEST       STRATUM PSU ;
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     ;
```

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

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 Year 2000 application (added 01/21/98)

The variance estimation methods of this document may be applied to the 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