SIMPLIFIED DESIGN STRUCTURES FOR NHANES I VARIANCE ESTIMATION

Michael Rowland, Van Parsons and Diane Makuc
National Center for Health Statistics
Michael Rowland, Rm 2-58, 3700 East-West Hwy, Hyattsville, MD 20782

KEY WORDS: survey, variance, design

Introduction

The first National Health and Nutrition Examination Survey (NHANES I) was conducted from 1971 through 1975 based on an approach of multistage, stratified, national probability samples of loose clusters of civilian noninstitutionalized persons. Users of data from the NHANES I were advised to take the complex sample design into account in analysis, because an assumption of simple random sampling may provide misleading results[1]. However, the design structure and documentation for NHANES I variance estimation have proven cumbersome to use. In particular, the use of strata and primary sampling units (PSU's) for variance estimation is not well documented for NHANES I data user tapes.

In order to provide data users with the capability of estimating variances which take into account the survey design, strata and primary sampling unit (PSU) codes have been provided on these tapes. However, current procedures require that users recode strata and PSU's to compute variances for some subsamples of the survey. This procedure has proven difficult for many users to implement.

This report documents simplified design structures for NHANES I variance estimation which will facilitate variance computation.

Survey Description

The National Health and Nutrition Examination Survey of 1971-75 (NHANES I) was the first in a series of Health and Nutrition Examination Surveys conducted by the National Center for Health Statistics. These surveys are unique in that the persons selected in the sample are both interviewed in the household and examined in mobile examination centers that are moved from one ,sampling location to another. The logistics of the survey require that the number of sampling locations be limited and a sufficient number of people examined at each site to achieve the desired number in the total sample. Therefore, the surveys are characterized by a relatively small number of PSU's with a relatively large number of sample persons in each PSU [2,3].

The NHANES I sample design was similar to those of subsequent National Health and Nutrition Examination Surveys (NHANES II and the Hispanic HHANES)[4,5]. All of the surveys have used complex, multistage, stratified, clustered samples of defined populations. In hierarchical order, the stages of selection for each survey were primary sampling units (PSU's), segments consisting of clusters of households, households within clusters, and eligible persons within households.

The NHANES I sample design has been described elsewhere[2,3], but aspects of the design pertaining to variance estimation are presented here. The survey was conducted at 100 locations across the United States from 1971 through 1975 and consists of 6 overlapping nationally representative samples as shown in table 1. The seventh sample referred to in the table is a composite of all persons aged 25-74 years examined in 1971-75. However, statistical weights are not available for this sample.

All 20,749 examined persons in the first 65 survey locations (or stands) received a specifically designed nutrition-related examination (1-65 nutrition). In addition, approximately 20 percent of those ages 25-74 years, or 3,854 persons, received a more detailed examination concerning other aspects of health and health care needs (1-65 detail). In order to produce national estimates of the nutritional status of the U.S. population at an early date, a probability subsample of 35 stands of the 65 stands were selected and 10,127 persons were examined in 1971-72 (1-35 nutrition). Approximately 20 percent of persons 25-74 years in the 35 stand sample or 1,892 persons also received a detailed examination (1-35 detail). To increase the size of the subsample of 25-74 year old adults, the design further provided for selection of an additional national sample of 35 survey locations, sometimes referred to as the Augmentation survey. In this sample, 3,059 adults were given a detailed examination in 1974-75 (66-100 detail). The first 65 location detailed sample combined with this additional 35-location sample form a 100-PSU national probability sample in which the combined number of persons is 6,913 (1-100 detail). The 100-PSU combined sample of nutrition and detailed persons aged 25-74 years is the baseline sample for the NHANES I Epidemiologic Followup Study (1-100 nutrition and detail).[6]

Variance Design Structure

The essential feature of the NHANES I for variance estimation is the selection of primary sampling units (PSU's) from strata.

NHANES I Variance Design Structure

The current NHANES I design summarized in table 2 for each of the subsamples described previously can be characterized as having the following structure:

1) 10 certainty strata with selections of segments designated as PSU's for variance calculations and with multiple PSU's (segments) per stratum,
2) 25 noncertainty strata with
   a) selection of 3 PSU's per stratum for survey locations 1-100;
   b) paired selections of PSU's per stratum for survey locations 1-65;
   c) selection of a single PSU per stratum for survey locations 1-35 and locations 66-100 (Augmentation sample). In the microdata tape documentation, NCHS has recommended that for these samples the 25 noncertainty strata be collapsed into 13 pseudo strata for variance computation.

In order to alleviate the need for data users to recode strata and PSU's for variance computation, NCHS is making available upon request a computer data tape with revised indexing of strata and PSU's. This revision also reflects a simplification of the variance design structure including the random allocation of multiple PSU's(segments) into 3 pseudo-PSU's per stratum for certainty strata.

Simplified NHANES I Variance Design Structure

The redefinition of NHANES I strata and PSU's is as follows:

1) For certainty strata 1-10, PSU's (segments) were randomly allocated into 3 pseudo PSU's per stratum.
2) For noncertainty strata 11-35 in the 65 stand and 100 stand subsamples, each PSU(stand) was assigned a code of 1,2,or 3 as follows:

Stands	New PSU code
1-35	1
36-65	2
66-100	3

To estimate variances in the 35 stand and Augmentation subsamples, the single-PSU strata have been grouped for noncertainty strata 11-35 using the collapsed strata technique described by Hansen, Hurwitz, and Madow[7]. Note that documentation for the original variance design is incomplete in that it leaves stratum 13 with a single unmatched PSU. In the new variance design, data from stratum 13 is combined with stratum 11 from the same geographic region, one of the original sample design stratification variables.

Two pairs of pseudo strata-PSU codes are now available on computer tape which reflect this revised indexing of strata and PSU's. The first pair should be used for NHANES I samples 1-35 detail, 1-35 nutrition and 66-100 detail. The second pair should be used with samples 1-65 detail, 1-65 nutrition, 1-100 detail and 1-100 nutrition and detail.

Evaluation

The effects of randomly assigning multiple "PSU's" (segments) to three pseudo-PSU's is investigated here by comparing design effects for estimates of proportions and means.

In survey research, the design effect is commonly defined to be the ratio of the actual variance estimate for a statistic taking into account the complex sample to the corresponding variance assuming a simple random sample. The design effect is used here to summarize conveniently the effects of the complex sample design on the precision of estimates from the survey data and as a means to compare the three variance design structures.

The SESUDAAN program of the Research Triangle Institute[8] was used to compute variances according to the Taylor series approach, and the BRR program of the National Center for Health Statistics[9] was used to compute variances according to the Balanced Repeated Replication approach. The design structure for the BRRP estimates is similar to that for the simplified variance design structure.

The effects of randomly assigning the multiple PSU's to three pseudo-PSU's was investigated by comparison of design effects for a mean and two percents. In addition to the two analytic variables mean body weight and percent overweight, percent born in July was chosen as a comparison variable under the rationale that it was uncorrelated with design variables. Design effects were calculated for 3-4 age groups and then averaged. As is shown in Tables 3 and 4, the average design effects were similar for the multiple(i.e. original) and two/three(i.e. simplified) PSU variance designs. The average design effects were generally similar for the BRRP approach and the two Taylor linerization approaches(i.e. original and simplified).

The BRRP average design effects for the 1-100 detail sample design for each of the three variables examined here were extremely large, averaging more than 3.0. The reason for this is unclear, and these atypical design effects have been excluded from the averages across samples in table 3.

The size of the average design effects shown in Table 3 and 4 tells you little about the distribution, however, and as one can see from Table 5 the distributions were somewhat more concentrated in the Taylor linearization approaches than in the BRRP approach. Overall, 65-70 percent of average design effects were in the range 1.0-1.99 for the Taylor linearization approaches compared to 50-63 percent for the BRRP approach.

Discussion

At least for the variables considered here, the random allocation of PSU's in the certainty strata to form a two/three PSU per stratum design has not substantially altered the design effects or the corresponding variances(not shown here).

The magnitude of the average design effects for these NHANES I variables is similar to those found for NHANES II variables in a study by Kovar and Johnson[10] using smaller age groups. However, the distribution of design effects is more concentrated in the range 1.0-1.99 for the NHANES II.

An additional observation for these NHANES I data is that the Taylor linearization approach appears to provide more consistent variance estimates, i.e. more concentrated range of design effects, than does the BRRP approach.

Both the BRRP and Taylor series expansion approaches to variance estimation are available in a number of statistical software packages that can incorporate the sampling weights and the design structure into the analysis. Landis[1] has found working with the NHANES I data that those packages using the BRRP approach benefit in cost and computing time efficiency with the simplified design structure allocation of PSU's (segments) in the certainty strata as was done here.

It is hoped that the simplified design structure for variance estimation presented here will facilitate the task of the NHANES I data analyst. The revised indexing of strata and PSU's is being made available on computer data tape by NCHS.

References

1. Landis, JR, JM Lepkowski, SA Eklund, and SA Stehouwer: A Statistical Methodology for Analyzing Data from a Complex Survey: The first National Health and Nutrition Examination Survey. Vital and Health Statistics. Series 2, No 92. DHHS Pub. No. (PHS)82-1366. Public Health Service. Washington.U.S. Government Printing Office.

2. National Center for Health Statistics: Plan and operation of the Health and Nutrition Examination Survey, United States, 1971-73. Vital and Health Statistics. Series 1-Nos. 10a and 10b. DHEW Pub. No. (HSM) 73-1310. Health Services and Mental Health Administration. Washington. U.S. Government Printing Office, Feb. 1973.

3. National Center for Health Statistics: Plan and operation of the HANES I Augmentation Survey of Adults 25-74 years, United States, 1974-75. Vital and Health Statistics. Series 1-No. 14. DHEW Pub. No. (PHS) 78-1314. Public Health Service. Washington. U.S. Government Printing Office, June 1978.

4. National Center for Health Statistics: Plan and Operation of the Second National Health and Nutrition Examination Survey, 1976-80. Vital and Health Statistics. Series 1 No 15. DHHS Pub. No. (PHS) 81-1317. Public Health Service. Washington. U.S. Government Printing Office.

5. National Center for Health Statistics: Maurer, KR and others: Plan and Operation of the Hispanic Health and Nutrition Examination Survey, 1982-84. Vital and Health Statistics. Series 1 No 19. DHHS Pub. No. (PHS) 85-1321. Public Health Service. Washington. U.S. Government Printing Office.

6. National Center for Health Statistics: Cohen, BB and others: Plan and Operation of the NHANES I Epidemiologic Followup Study: 1982-84. Vital and Health Statistics. Series 1, No. 22. DHHS Pub. No.(PHS) 87-1324. Public Health Service. Washington. U.S. Government Printing Office, June 1987.

7. Hansen, MH, WN Hurwitz, and WG Madow: Sample Survey Methods and Theory. Vol. 1. New York. John Wiley & Sons, Inc., 1953.

8. Shah, BV: SESUDAAN: Standard Errors Program for Computing of Standardized Rates from Sample Survey Data. RTI/5250/00-01S. Research Triange Institute, Research Triange Park, NC, 1981.

9. Jones, G. The NCHS BRR program.

10. Kovar, MG and C Johnson: Design Effects from the Mexican American Portion of the Hispanic Health and Nutrition Examination Survey: a Strategy for Analysts, Proceedings of the Survey Research Methods Section, American Statistical Association, 1986.

Table 1. Number of survey locations, type of examination, years of data collection, age of target population, and number of examined persons for persons for NHANES I data.

Survey locations and examination in sample design	Year	Age in years of target population	Number of examined persons
1-35 detail	1971-72	25-74	1,892
1-35 nutrition	1971-72	1-74	10,127
1-65 detail	1971-74	25-74	3,854
1-65 nutrition	1971-74	1-74	20,749
66-100 detail[1]	1974-75	25-74	3,059
1-100 detail[2]	1971-75	25-74	6,913
1-100 nutrition and detail[2,3]	1971-75	25-74	14,407

[1] Augmentation

[2] Includes augmentation sample

[3] Sample for the NHANES I Epidemiologic Followup Study.

Table 2. Number of primary sampling units (PSU's) by stratum number for three survey portions of the NHANES I design: United States, 1971-75

Stratum number	Total Stands 1-100 (1971-75)	Survey Portion Stands 1-35 (1971-72)	Stands 36-65 (1972-74)	Stands 66-100 (1974-75)
Total	1499	959	304	236
1-10	1424	934	279	211
1	190	105	64	21
2	123	67	39	17
3	143	82	43	18
4	177	97	59	21
5	221	123	74	24
6	105	83	0	22
7	131	108	0	23
8	82	61	0	21
9	110	89	0	21
10	142	119	0	23
11-35	75	25	25	25
11	3	1	1	1
12	3	1	1	1
13	3	1	1	1
14	3	1	1	1
15	3	1	1	1
16	3	1	1	1
17	3	1	1	1
18	3	1	1	1
19	3	1	1	1
20	3	1	1	1
21	3	1	1	1
22	3	1	1	1
23	3	1	1	1
24	3	1	1	1
25	3	1	1	1
26	3	1	1	1
27	3	1	1	1
28	3	1	1	1
29	3	1	1	1
30	3	1	1	1
31	3	1	1	1
32	3	1	1	1
33	3	1	1	1
34	3	1	1	1
35	3	1	1	1

NOTE: In the certainty strata 1-10, PSU's are segments(i.e. a cluster of households). In the noncertainty strata 11-35, PSU's are counties or groups of contiguous counties.

Table 3. Average design effects for selected variables according to variance estimation procedure and sex.
Six NHANES I national samples combined.

		Total	Males	Females
Mean weight	BRRP	1.2	1.7	1.3
	Taylor linearization			
	Simplified design	1.7	1.7	1.3
	Original design	1.6	1.7	1.4
Percent overweight	BRRP	1.7	1.8	1.8
	Taylor linearization			
	Simplified design	1.6	1.6	1.5
	Original design	1.6	1.6	1.5
Percent born in July	BRRP	1.8	1.9	1.8
	Taylor linearization			
	Simplified design	1.6	1.6	1.7
	Original design	1.6	1.7	1.6

Averages of age groups.
Average design effects for the BRRP procedure do not include the 1-100 stand sample.

Table 4. Average design effects for selected variables according to variance estimation procedure and sample.

		1-35 D	1-35 N	1-65 D	Sample 1-65 N	66-100 D	1-100 D
Weight	BRRP	1.8	1.5	1.5	1.3	1.2	3.3
	Taylor linearization						
	Simplified design	1.7	1.7	1.5	1.5	1.2	1.6
	Original design	1.7	1.7	1.5	1.5	1.2	1.6
Overweight	BRRP	1.4	2.1	1.6	2.0	1.4	3.1
	Taylor linearization						
	Simplified design	1.5	2.0	1.5	1.7	1.1	1.5
	Original design	1.3	1.9	1.6	1.7	1.4	1.5
Born in July	BRRP	2.1	1.7	1.7	1.9	1.8	3.2
	Taylor linearization						
	Simplified design	1.6	1.7	1.5	1.9	1.4	1.6
	Original design	1.8	1.6	1.5	1.8	1.6	1.6

Averages of age-sex specific groups.

Table 5. Distribution of age-sex specific design effects for selected variables according to variance estimation procedure.
Six NHANES I national samples combined.

	Weight			Overweight			Born in July		
Design Effect	BRRP	Taylorization Simplified design	Original design	BRRP	Taylorization Simplified design	Original design	BRRP	Taylorization Simplified design	Original design
					Number of design effects				
Total	34	40	40	30	36	36	34	40	40
<0.5	2	1	1						1
0.5-0.99	7	5	5		5	4	3	4	5
1.0-1.49	10	12	13	12	15	13	8	14	8
1.5-1.99	7	16	14	7	9	12	10	12	18
2.0-2.49	4	5	4	7	6	6	7	6	5
2.5-2.99	4	1	3	4		1	5	3	2
3.0-3.49							1	1	1
3.5-3.59					1				
4.0-4.49									
4.5-4.99									
5.0+									

Distributions of age-sex specific design effects for BRRP procedure do not include the 1-100 stand detail sample.