SIMPLIFIED VARIANCE ESTIMATION FOR THE NATIONAL HEALTH INTERVIEW SURVEY

Robert J. Casady, Bureau of Labor Statistics, Van L. Parsons and
Cecelia B. Snowden, National Center for Health Statistics

1. Introduction

The survey design and methodology for first order estimation (i.e., aggregates, means, proportions) for the National Health Interview Survey (NHIS) are very complex. These complexities result from efforts by statisticians at the U.S. Census Bureau and the National Center for Health Statistics (NCHS) to produce highly reliable estimates of health related characteristics of the United States population under rigorous cost constraints. Unfortunately, the complex sample design and estimation procedures utilized in the NHIS make the direct estimation of the variance of the first order estimators virtually impossible, so that other indirect methods such as the Jackknife, Balanced Repeated Replication (BRR), linearization or bootstrapping must be used. Rust (1985) notes that at present, on theoretical grounds, there appears to be little to choose among these various methods and in most cases it is the availability, cost and ease of use of computer software that are the main considerations in the practical choice among these methods.

Historically, the BRR procedure has been used for NHIS variance estimation but concurrent with implementation of the new design in 1985, it was decided to change to the linearization procedure. There were three primary reasons for this decision. First, the new design called for two primary sampling units (PSUs) to be selected PPS without replacement from each non-self-representing stratum, so it was no longer reasonable to assume that first stage selection was independent. This lack of independence would complicate the BRR procedure and require major modification of the existing NHIS variance estimation software. Secondly, NHIS estimation procedures are based on a sequence of ratio adjustments so that the computation of linearized estimates of variance is relatively simple and hence, inexpensive in terms of computer storage and CPU time compared to the alternative replicated procedures. Finally, it was felt that the linearization procedure would give statisticians at the Census Bureau and NCHS greater latitude and flexibility in studying operational and statistical aspects of the NHIS survey design and estimation methodology (i.e., estimation of variance components, variance reduction due to ratio adjustment, etc.). The results presented in this paper and in Parsons and Casady (1986) demonstrate the power and flexibility of the new NHIS variance estimation software in evaluating NHIS estimation procedures.

In the next section, the model for the 1985 NHIS sample design and general NHIS estimation methodology will be briefly reviewed and the development of the NHIS linearized variance will be discussed. In the third section, an approximation to the linearized form of the NHIS estimator will be suggested and the resulting estimator of variance will be empirically compared to the variance estimator developed in Section 2. In Section 4, the use of SESUDAAN to estimate variances for NHIS will be proposed. The implicit model assumptions required to use the SESUDAAN software will be briefly discussed and an empirical comparison of estimates from SESUDAAN and the linearized variance estimator will be presented. In Section 5, the results of the empirical studies will first be summarized and then used as the basis for recommendations regarding variance estimation methodology for NHIS.

2. NHIS Design and First Order Estimation

Prior to specifying the NHIS design and estimation models, the linearization variance estimation technique will be briefly reviewed. For a specified sample design, say D, let $\hat{V}(D,\hat{L})$ be an estimator for the variance of the simple linear estimator

$$\hat{L} = \sum_{i=1}^{N} a_i X_i \text{ , where}$$

$$a_i = \begin{cases} 1 & \text{if the } i^{th} \text{ frame element is selected} \\ 0 & \text{otherwise} \end{cases}$$

The distribution of the random vector $\underline{a} = (a_1, a_2 \ldots, a_N)$ is determined by the design D and X_i is a known real number associated with the i^{th} frame element. Assume, via a Taylor's series expansion, the linear approximation to the non-linear estimator $\hat{\theta}$ is given by

$$\hat{L}(\hat{\theta}) = \sum_{i=1}^{N} a_i X_i^*.$$

The linearized variance estimator for $\hat{\theta}$ is then given by $\hat{V}(D, \hat{L}(\hat{\theta}))$. It is important to note that the algebraic form of the non-linear estimator $\hat{\theta}$ determines the formula required to calculate the real numbers X_i^*. An excellent discussion of the linearization estimation procedure can be found in Woodruff (1971).

For the purposes of this paper, it is sufficient to know that NHIS utilizes a stratified multistage cluster design with PPS sampling at the first stage and (assumed) simple random sampling at all higher stages. A more detailed description of the design and the algebraic formula for $\hat{V}(D,\hat{L})$ can be found in Parsons and Casady (1986).

The general form of the NHIS estimators for person based aggregates and means (or proportions) are

$$X' = \sum_{i=1}^{N} a_i \hat{\lambda}_{2i} \hat{\lambda}_{1i} [W_i X_i Y_i] \qquad (1)$$

$$\text{and } \bar{X}' = \frac{\sum_{i=1}^{N} a_i \hat{\lambda}_{2i} \hat{\lambda}_{1i} [W_i X_i Y_i]}{\sum_{i=1}^{N} a_i \hat{\lambda}_{2i} \hat{\lambda}_{1i} [W_i Y_i]} \qquad (2)$$

where for the i^{th} person, W_i is the simple inflation weight (i.e., the inverse of the probability of selection), X_i is the value of the variate of interest, Y_i is a variate indicating domain of study membership, $\hat{\lambda}_{1i}$ is the first stage ratio adjustment factor and $\hat{\lambda}_{2i}$ is the second stage ratio (or post-stratification) adjustment factor. For the purpose of this paper, bracketed quantities in first order estimators are to be considered fixed real constants.

As both the first and second stage adjustment factors depend on the random vector $\underline{a}$, it is clear that X' and $\bar{X}'$ are non-linear estimators. Actually the estimators X' and $\bar{X}'$ are even more complex than the representations in (1) and (2) indicate because the second stage ratio adjustment factors are functionally dependent on the first stage ratio adjustment factors. Detailed representations for NHIS estimators may be found in Schaible (1975). Parsons and Casady (1986) derive the linearized forms of X' and $\bar{X}'$ and in addition, Parsons has written a computer software package to produce variance estimates utilizing the estimators $\hat{V}(D,\hat{L}(X'))$ and $\hat{V}(D,\hat{L}(\bar{X}'))$.

3. A Variance Estimator Using a Simplified First Order Estimation

The complex algebraic structure of the NHIS estimators X' and $\bar{X}'$ necessitates a computer program that requires a relatively large amount of internal storage and CPU time to produce variance estimates using $\hat{V}(D,\hat{L}(X'))$ and $\hat{V}(D,\hat{L}(\bar{X}'))$. Although this is not a problem if only a few variance estimates are required, such is not the case for large production runs. One possible approach to reducing storage and CPU requirements is to simplify the estimation equations by assuming a simpler algebraic form for X' and $\bar{X}'$.

Motivated by the fact that the first stage ratio adjustment factors are only applied to the inflation weights of persons in non-self-representing PSUs and the fact that results in Parsons and Casady (1986) indicated the first stage ratio adjustment had little impact on the variance of the first order estimators, it was decided to investigate the performance of the linearization variance estimator when it was assumed that

$$X' = \sum_{i=1}^{N} a_i \hat{\lambda}_{2i} [\hat{\lambda}_{1i} W_i X_i Y_i] \qquad (3)$$

$$\text{and } \bar{X}' = \frac{\sum_{i=1}^{N} a_i \hat{\lambda}_{2i} [\hat{\lambda}_{1i} W_i X_i Y_i]}{\sum_{i=1}^{N} a_i \hat{\lambda}_{2i} [\hat{\lambda}_{1i} W_i Y_i]} \qquad (4)$$

That is, the linear approximations, say $\hat{L}_1(X')$ and $\hat{L}_1(\bar{X}')$, were derived under the assumption that the first stage ratio adjustment factors were not random variables.

This simplifying assumption had the effect of reducing the CPU time by a factor of 20%, and internal storage by a factor of 40%. To evaluate the performance of the simplified estimation procedure, relative to the standard procedure, the ratio of estimated standard errors $[\hat{V}(D,\hat{L}_1(\cdot)) / \hat{V}(D,\hat{L}(\cdot))]^{\frac{1}{2}}$, was calculated for a broad range of NHIS variates (including demographic, socio-economic and health variates) and broad range of domains of study. Some of the results of this study are presented in Tables 1, 2 and 3. These results indicate that the simplified estimator produces a close approximation to the estimate of standard error provided by the standard linearized estimator. In general, it appears to provide a very slight over-estimate of standard error. However, the largest observed over-estimate was only 3.05% larger than standard linearized estimate (for health characteristics the largest over-estimate was only 1.70% larger). All observed under-estimates were less than 2% smaller than the standard linearized estimate and nearly all observed under-estimates were less than 1% smaller (only four were between 1% and 2% smaller). The probable reason for the simplified estimator to produce slightly larger estimates than the standard estimator is that the simplified estimator ignores the slight variance reducing property of the first stage ratio adjustment factors.

4. SESUDAAN/RATIOEST Estimates of Variance

The results of the preceeding section provide some justification for the use of a simplified linear form when the objective is to "mass produce" variance estimates for NHIS. Unfortunately, this information is of little value to the consumer of NHIS public use tapes as the NHIS linearization program is not available to the general public. Recently Cohen, Burt and Jones (1986) studied several variance estimation programs including the widely available SESUDAAN/RATIOEST, SUPERCARP and PSALMS programs. These three programs utilize essentially identical linearization procedures and hence, produce equivalent estimates of standard errors (Francis and Sedransk 1979). However, Cohen et al recommended the use of the SESUDAAN/RATIOEST program when the desired output was estimates of population means and totals and their associated standard errors in a cross-tabluated format.

Their recommendation was based primarily on programming facility and CPU time.

Unfortunately, the sample design and estimators assumed in the SESUDAAN/RATIOEST program do not completely coincide with the sample design and estimators utilized by NHIS. Specifically, the SESUDAAN/RATIOEST design, denoted D*, assumes simple random sampling without replacement in the first stage where as NHIS utilizes PPS sampling without replacement. Further SESUDAAN/RATIOEST assumes (unless the user does some "fixing up") simple linear estimators or ratios of simple linear estimators. Thus, using SESUDAAN/RATIOEST for NHIS variance estimation requires that the user implicitly assume the design D* and estimators of the form

$$X' = \sum_{i=1}^{N} a_i \left[\hat{\lambda}_{2i}\, \hat{\lambda}_{1i}\, W_i\, X_i\, Y_i\right] \qquad (5)$$

and

$$\bar{X}' = \frac{\sum_{i=1}^{N} a_i \left[\hat{\lambda}_{2i}\, \hat{\lambda}_{1i}\, W_i\, X_i\, Y_i\right]}{\sum_{i=1}^{N} a_i \left[\hat{\lambda}_{2i}\, \hat{\lambda}_{1i}\, W_i\, Y_i\right]} \qquad (6)$$

It is not really possible to analytically forecast the impact of these assumptions on the variance estimators (Casady 1985). We were reasonably confident that the use of the design D* instead of D would have relatively little impact because even though the PSU's were selected PPS, the size measures were relatively homogeneous within stratum. However, we also suspected that absorbing the first and second stage adjustment factors into the base weights would tend to produce conservative estimates.

Letting $\hat{L}_2(\cdot)$ represent the linear approximation to estimators of the form assumed in (5) and (6), SESUDAAN/RATIOEST was used to produce estimates of variance $\hat{V}(D^*, \hat{L}_2(\cdot))$ for the same NHIS variates and domains of study as in the previous section. As in the preceding section, the ratios of $[(\hat{V}(D^*, \hat{L}_2(\cdot)) / \hat{V}(D, \hat{L}(\cdot)))]^{1/2}$ were calculated and are presented in Tables 4, 5 and 6.

Based on the preceeding discussion, the results are not surprising. In general, the standard error estimates produced by SESUDAAN are 10 to 15 percent larger than the estimates using the standard linearization estimator for health characteristics and socio-economic characteristics. As noted in the previous section, this tendency to produce conservative estimates is undoubtably due to the fact that the variance reducing property of the ratio adjustment factors is ignored. This point is further illustrated by the fact that the SESUDAAN/RATIOEST estimates of standard error for demographic attributes, which are highly correlated with the poststratification attributes, are commonly more than 25% larger than $[\hat{V}(D, \hat{L}(\cdot))]^{1/2}$.

5. Summary and Recommendation

It has been demonstrated that, for a wide range of health variates and domains of study, the simplified linearized form $\hat{L}_1(\cdot)$ produces estimates of standard error that very closely approximate to those produced by the standard linearized form $\hat{L}(\cdot)$. The linearized estimator $\hat{L}_2(\cdot)$ (together with the design D*), produce estimates of standard error that are 10 to 15 percent larger than those produced by $\hat{L}(\cdot)$ (under the design D). Thus, we feel that we can strongly recommend that $V(D, \hat{L}_1(\cdot))$ be used when larger numbers of estimates are required. Further, we feel that $V(D^*, \hat{L}_2(\cdot))$ can only be marginally recommended as an acceptable procedure for estimating standard errors of health variates. Estimates under this procedure will tend to be conservative in the range of 10 to 15 percent. One word of caution is necessary. The estimator $\hat{V}(D^*, \hat{L}_2(\cdot))$ should not be used to estimate standard error for estimates of population totals (or proportions) that closely correspond to post-stratification domains, as the estimates of standard error will tend to be much too large.

Although the results presented in this paper cannot be extended to other complex sample surveys without extreme caution, they do serve to illustrate a general short coming of existing, supported software programs for complex surveys. Specifically, the existing software programs assume simpler sample designs and estimation procedures than are commonly used in the real world. Consequently, as we have seen in this paper, variance estimates produced by these programs will, in general, be conservative because the increased precision of the complex design and estimation methodology is not accounted for by the variance estimator.

REFERENCES

Casady, Robert J. (1985), "Variance Estimates From the SESUDAAN Program", National Center for Health Statistics Internal Memorandum.

Cohen, Steven B., Burt, Vicki L. and Jones, Gretchen K. (1986), "Efficiencies in Variance Estimation for Complex Survey Data", The American Statistician, Vol. 40, No. 2, 157-164.

Francis, I. and Sedransk, J. (1979), "A Comparison of Software for Processing and Analyzing Survey Data", Bulletin of the International Statistical Institute, 48, 1-31.

Parsons, Van L. and Casady, Robert J. (1986), "Variance Estimation and the Redesigned National Health Interview Survey", Proceedings of the American Statistical Association, Section on Survey Research Methods, in press.

Rust, Keith (1985), "Variance Estimation for Complex Estimators in Sample Surveys", Journal of Official Statistics, Vol. 1, No. 4, 381-397.

Schaible, Wesley L. (1975), "Basic Estimation of NHIS; 1973 Design", National Center for Health Statistics Internal Memorandum.

Shah, B.V. (1981b), "Standard Errors Program for Computing Standardized Rates for Sample Survey Data," technical report, Research Triangle Park, NC: Research Triangle Institute.

Woodruff, R.S. (1971), "A Simple Method for Approximating the Variance of a Complicated Estimate", Journal of the American Statistical Association, 66, 411-414.

Table 1 - Observed values of $[\hat{V}(D,\hat{L}_1,(\hat{p})) / \hat{V}(D,\hat{L}(\hat{p}))]^{1/2}$ where $\hat{p}$ is the NHIS estimate of the proportion of the population in the domain of study with a specified demographic attribute.

Domain of Study	Demographic Attribute					
			Age			
	Race (Black)	Sex (Male)	≤ 17	18-44	45-64	65 +
N.E.	1.0019	1.0300	1.0146	1.0202	.9973	.9981
M.W.	1.0305	1.0000	1.0000	1.0161	.9972	1.0040
South	.9979	1.0132	1.0000	1.0063	1.0000	.9975
West	.9939	1.0000	.9973	.9955	1.0022	.9986
Poverty	.9963	.9947	1.0033	1.0069	1.0049	1.0054
Females South	.9862	-	1.0042	1.0000	.9935	.9977
Blacks South	-	1.0059	1.0000	1.0085	1.0018	.9925

Table 2 - Observed values of $[\hat{V}(D,\hat{L}_1,(\hat{p})) / \hat{V}(D,\hat{L}(\hat{p}))]^{1/2}$ where $\hat{p}$ is the NHIS estimate of the proportion of the population in the domain of study with a specified socio-economic attribute.

Domain of Study	Socio-Economic Attribute				
			Household Income		
	Below Poverty	<10,000	10,000-20,000	20,000-35,000	35,000 +
U.S.	.9954	.9973	1.0071	1.0041	1.0030
N.E.	.9863	.9868	1.0170	1.0063	1.0025
M.W.	1.0166	1.0142	1.0041	1.0022	1.0045
South	.9908	1.0000	.9978	1.0076	.9967
West	.9992	1.0125	1.0000	1.0000	1.0029
Females	NC	.9972	1.0034	1.0039	1.0028
Blacks	NC	1.0000	1.0141	1.0115	1.0016
Age 65 +	1.0037	NC	NC	NC	NC
Females South	NC	.9948	1.0000	1.0049	.9984
Blacks South	NC	1.0012	1.0230	1.0193	.9916

NC - Not Calculated

Table 3 - Observed values of $[\hat{V}(D,\hat{L}_1,(\hat{p})) / \hat{V}(D,\hat{L}(\hat{p}))]^{1/2}$ where $\hat{p}$ is the NHIS estimate of the proportion of the population in the domain of study with a specified health characteristic.

Domain of Study	Health Characteristic							
	Health Status					One or More Bed Days	One or More Doctor Visits	One or More Hospital Days
	Excellent	Very Good	Good	Fair	Poor			
U.S.	1.0000	.9939	.9950	1.0034	1.0086	1.0096	1.0000	1.0000
N.E.	1.0000	.9942	1.0023	1.0170	1.0136	.9929	1.0027	1.0055
M.W.	1.0149	.9965	1.0074	1.0052	1.0058	1.0102	.9972	1.0000
South	.9966	1.0000	1.0000	1.0000	.9988	1.0000	.9972	1.0000
West	1.0000	1.0000	.9972	1.0015	1.0018	.9955	1.0000	.9981
Females	1.0000	1.0000	1.0000	1.0030	1.0051	1.0093	1.0000	1.0000
Blacks	.9941	1.0000	.9956	.9953	.9978	1.0101	1.0134	1.0034
Age 65 +	.9982	.9956	1.0065	1.0000	1.0077	1.0000	.9974	1.0056
Below Poverty	.9956	1.0000	1.0000	1.0000	1.0000	.9958	1.0020	1.0071
Female South	.9972	1.0000	1.0000	.9965	.9944	1.0000	1.0000	.9977
Blacks South	.9932	.9987	.9948	.9946	.9984	1.0047	1.0052	1.0025

Table 4 - Observed values of $[\hat{V}(D,\hat{L}_1,(\hat{p})) / \hat{V}(D,\hat{L}(\hat{p}))]^{1/2}$ where $\hat{p}$ is the NHIS estimate of the proportion of the population in the domain of study with a specified demographic attribute.

Domain of Study	Demographic Attribute					
			Age			
	Race (Black)	Sex (Male)	$\leq$ 17	18-44	45-64	65 +
N.E.	1.3190	1.1667	1.1744	1.1197	1.0405	.9839
M.W.	1.4432	1.1915	1.1250	1.1772	1.1940	1.1897
South	1.7857	1.3784	1.1964	1.3676	1.3137	1.3061
West	1.1446	1.1356	1.0784	1.0761	1.1522	1.1781
Poverty	1.4177	1.1963	1.3238	1.3136	1.1045	1.2051
Females South	1.6957	-	1.1935	1.3239	1.2500	1.3621
Blacks South	-	1.4557	1.3636	1.4021	1.500	1.8462

Table 5 - Observed values of $[\hat{V}(D,\hat{L}_1,(\hat{p})) / \hat{V}(D,\hat{L}(\hat{p}))]^{\frac{1}{2}}$ where $\hat{p}$ is the NHIS estimate of the proportion of the population in the domain of study with a specified socio-economic attribute.

Domain of Study	Socio-Economic Attribute				
			Household Income		
	Below Poverty	<10,000	10,000-20,000	20,000-35,000	35,000 +
U.S.	1.1091	1.1132	1.0526	.9429	1.0123
N.E.	1.0076	1.0000	.9111	.6732	.9016
M.W.	1.0000	1.0088	1.1154	1.0388	1.0316
South	1.1474	1.1188	1.0213	1.0777	1.0227
West	1.0305	1.0380	1.0537	1.0268	1.0152
Females	NC	1.1017	1.0161	.9577	1.0000
Blacks	NC	1.1105	1.0400	1.1325	1.0299
Age 65 +	1.1325	NC	NC	NC	NC
Females South	NC	1.0952	.9725	1.0481	1.0154
Blacks South	NC	1.1492	1.0330	1.1535	.9942

NC - Not Calculated

Table 6 - Observed values of $[\hat{V}(D,\hat{L}_1,(\hat{p})) / \hat{V}(D,\hat{L}(\hat{p}))]^{\frac{1}{2}}$ where $\hat{p}$ is the NHIS estimate of the proportion of the population in the domain of study with a specified health characteristic.

Domain of Study	Health Characteristic							
	Health Status					One or More Bed Days	One or More Doctor Visits	One or More Hospital Days
	Excellent	Very Good	Good	Fair	Poor			
U.S.	1.1053	1.0227	1.1064	1.1364	1.0769	1.0612	1.0811	1.1500
N.E.	.9549	1.0430	1.0000	.9074	1.0000	.8480	.9639	1.0444
M.W.	1.1250	1.0500	1.0851	1.1667	1.0952	1.0737	1.0758	1.1316
South	1.0865	1.0000	1.1707	1.0930	1.0000	1.1711	1.1642	1.1081
West	1.1077	1.0104	1.0638	1.0889	1.0000	1.1017	1.1169	1.1250
Females	1.0781	1.0000	1.0893	1.0690	1.0588	1.0926	1.0714	1.1000
Blacks	1.0956	1.0317	1.0786	1.1194	1.0488	1.0569	1.1604	1.0926
Age 65 +	1.0345	1.0220	1.0204	1.0103	1.0806	1.0103	1.0241	1.0270
Below Poverty	1.0902	1.0000	1.1360	1.1013	1.0192	1.1504	1.0755	1.0635
Female South	1.0496	1.0108	1.1262	1.0364	.9737	1.1477	1.1507	1.0769
Blacks South	1.1105	1.0275	1.0883	1.1078	.9831	1.0127	1.1288	1.0921