STRATUM PARTITION, COLLAPSE AND MIXING IN CONSTRUCTION OF BALANCED REPEATED REPLICATION VARIANCE ESTIMATORS

Van L. Parsons, National Center for Health Statistics
John L. Eltinge , Texas A&M University and Bureau of Labor Statistics
Van L. Parsons, 6525 Belcrest Road, room 915, Hyattsville, MD 20782, (vlp1@cdc.gov)

Key Words: Cluster Sample, Identification Risk, Complex Survey Design, National Health Interview Survey (NHIS), Primary Sample Unit (PSU), Stratified Multistage Sample Survey.

1 Introduction

Since 1957 the National Health Interview Survey (NHIS) has been a large-scale governmental survey fielded to assess the health status of the U.S. noninstitutionalized civilian population. This survey, sponsored by the National Center for Health Statistics (NCHS), primarily collects and disseminates self-reported categorical health statistics and demographic information. Historically, the focus of the NHIS has been to produce accurate and reliable statistics at the national level. In recent times state public health officials and non-government research organizations have had increased interest in using the very rich NHIS database to study specific states. Because of this interest, the NHIS redesign for 1995-2004 used a state-level stratification to accommodate possible state-level analyses. Past NHIS surveys had less potential for state-level estimation since they had used regional stratifications which introduced more variability into state-level design-based analyses.

The release of state-level data in public-use files poses a challenge to NCHS. First, it is felt that a state level file should be released in a way that allows reasonably efficient design-based analyses. This requires that NCHS release design structures having sufficient detail at the state level, e.g., strata, primary sample units (PSUs), weights or related replicate weights. However, it is imperative that release of this design information be done in a way that is consistent with NCHS regulations imposed by the Public Health Services Act, Section 308(d), requiring that publicly released data must avoid identification and disclosure risk.

The nature of the NHIS is such that the design levels are geographically clustered, and thus, identification risk increases when this information is enhanced by inclusion of a state identifier. To lessen this potential risk, a strategy for public data release should attempt to:

a. Reduce the released design information in a way that still permits accurate, stable variance estimation.

b. Ensure that the released design is compatible with existing complex-survey software, thus making the NHIS data of practical inferential value for a diverse group of data users.

The creation of a *2-PSUs-per-stratum*, balanced repeated replication (BRR) structure is a practical and reasonable method to address goals *(a)* and *(b)* above. In the next sections we discuss the creation of this structure and a means of assessment.

2 NHIS: State-Level Sample Design

The NHIS is intended to produce unbiased estimators at the state level, but it is not designed to produce reliable state estimates for all states. For pure design-based analytical strategies, about 10-15 of the larger states would support a stable linearization or replication approach to variance estimation; such states are the focus of this research. Most larger states tend to have 65% or more of their population concentrated in metropolitan areas. These areas are typically designated as self-representing (SR), and the sampling is somewhat dispersed throughout the entire area through the selection of a moderate to large number of geographically based clusters of housing units. The remaining areas are partitioned into non-self-representing (NSR) strata, and for these strata two primary sampling units (PSUs) consisting of counties or aggregates of counties are first selected to represent the stratum. For public data the distinct first-stage probabilities and joint probabilities of selection cannot be released, and thus, we will treat NSR PSUs as selected 2-per stratum with replacement. This is consistent with the

Table 1: Minority Density Strata

substratum	% black interval	% Hispanic interval
1	[0,10)	[0,5)
2	[0,10)	[5,10)
3	[0,10)	[10,30)
4	[0,10)	[30,60)
5	[0,10)	[60,100]
6	[10,30)	[0,5)
7	[10,30)	[5,10)
8	[10,30)	[10,30)
9	[10,30)	[30,60)
10	[10,30)	[60,100]
11	[30,60)	[0,5)
12	[30,60)	[5,10)
13	[30,60)	[10,30)
14	[30,60)	[30,60)
15	[30,60)	[60,100]
16	[60,100]	[0,5)
17	[60,100]	[5,10)
18	[60,100]	[10,30)
19	[60,100]	[30,60)
20	[60,100]	[60,100]
21	New Construction	

policy for nationally released data, and tends to produce conservative variance estimators (with respect to the design) over the NSR strata.

The remainder of this paper will focus on the SR strata with the emphasis on strategy points *(a)* and *(b)* above. Some additional comments about the combination of SR and NSR strata within a state appear at the end of section 3.

A thorough discussion of the NHIS design structures appears in NCHS (1999) and NCHS (2000). For completeness, a brief outline of the sampling structure within a self-representing stratum is now provided.

1. An SR stratum is partitioned into 1 up to 21 *minority density strata* defined by the percentage of black and Hispanic residents at the Census-Block (geographical clusters of housing units) level. Table 1 presents the minority density strata.

2. Within each density stratum of Table 1 the blocks are sorted by a Census Bureau ordering. (Details are given in Bureau of Census (1977) and other Census documents)

3. Within each stratum a systematic sample using a single random start is used to systematically select *block-strings* which are consecutive blocks in the sorted universe. Different sampling rates are targeted for the density strata of Table 1, and the initially selected *block-strings* are subsampled to meet targeted sampling rates. Next, households within sampled *block-strings* are sampled/partitioned to provide annual sample for the 10 years 1995-2004 of the NHIS.

4. The processes discussed in steps 2 and 3 are quite complicated. To simplify we conceptualize these selected *block-strings* as independent first-stage sampling units within a given density stratum (cf. *random order* conditions on systematic sampling, e.g., Wolter (1985) Chapter 7). Furthermore, we assume that the sampling and weighting is done is such a way that the usual Horvitz-Thompson estimator of total for any *block-string* is an unbiased estimator.

3 Construction of Design Information for State-Level Public-Use Variance Estimation

Before a *2-PSUs-per-stratum* public-use design approximation is constructed, we must have an implementable design structure that captures as much of the original sampling design as possible. This "best" conceptual design will be used as a baseline for comparisons, but generally involves information that cannot be included in a public-use data release for the confidentiality reasons reviewed in Section 1. Given the limitation in available universe information, our conceptual assumptions of step 4 above suggest that the "best" design will use an S^2-type variance estimator based upon estimated *block-string* totals. Now, since some density strata have only 1 selected *block-string*, standard methods are used to collapse *singleton* density strata with others, e.g., collapse population density strata based upon similar values for race/ethnic characteristics. The baseline conceptual design is the original design but supplemented with a moderate amount of collapse. This is demonstrated for an original SR stratum in Table 2 by columns 1 and 2. In this stratum the original density strata 3, 11, 9, 17, and 19 had just one unit each. As a result of this collapsing, we now have a design structure consistent with methods for producing standard linearization-based variance estimators, i.e,

For the moderately collapsed strata, $h = 1, 2, \ldots, L$, we have the point estimators

$\hat{Y} = \sum_{h=1}^{L} \hat{Y}_h$,

$\hat{Y}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} \hat{Y}_{hi}$,

$\hat{Y}_{hi}$ = Estimator of population total for density stratum h based on data from *block-string* i,

n_h the number of *block-strings*,

and we have a variance estimator of the form

$\hat{V}(\hat{Y}) = \sum_{h=1}^{L} \frac{1}{n_h(n_h-1)} \sum_{i=1}^{n_h} (\hat{Y}_{hi} - \hat{Y}_h)^2$

The above method is deficient for public release since the design structure uses the original (or masked) stratum and *block-string* labels which may result in possible identification risk. Furthermore, for most density strata $n_h > 2$, a structure not compatible with a *2-PSU-per-Stratum* design used with customary BRR variance estimators.

To meet objectives and *(a)* and *(b)* we will refine the initial design structure above to obtain

1. A *2-pseudo-PSU-per-Stratum* structure to be used with standard BRR software.

2. A large or moderate number of *degrees of freedom* $\approx$ (#PSUs - #Strata) computed according to an approximate design (under roughly equal pseudo-PSU sizes) to ensure reasonable variance estimator stability.

3. A reduced risk of identifying the geographical locations of original sampled *block-strings*.

To achieve this we will use the techniques of collapse, partition and mixing of density strata. First, we use the techniques of stratum collapse, and stratum partition to form roughly equal sized (with respect to the number of *block-strings*) pseudo-strata. Second, we mix the *block-strings* of the pseudo-strata to form pseudo-PSUs.

For the current application suppose that we target a pseudo-PSU to be an aggregate of about k original *block-strings*. Then the process of stratum collapse and partition uses the following steps.

C.1 Collapse the smaller density strata with "similar" strata to yield roughly equal sized, say $2k$, *block-strings* per pseudo-stratum.

C.2 Randomly partition large density strata or large collapsed strata of [C.1] into several pseudo-strata. Specifically, suppose that an original density stratum h is assigned to a stratum that will be partitioned into m pseudo-strata. Then using a method that achieves approximate balance, randomly assign n_h/m strings to each new pseudo-stratum.

Table 2: Collapsing and Partitioning of Density Strata

Original Density Strata[a]	Original n_h	New pseudo Density Stratum	New n_h
1	16	1a	10[b]
21	4	1b	10
2,*3*	3,*1*	2a	10
6,*11*	3,*1*	2a	.
7	2	2a	.
8,*9*	2,*1*	3a	12
14	2	3a	.
16,*17*	3,*1*	3a	.
18,*19*	2,*1*	3a	.

a: density stratum in italics has just 1 unit and is collapsed.
b: 8 units from stratum 1 and 2 units from stratum 21 are placed in 1a.

Table 2 illustrates the above steps C.1 and C.2. As an example, we see original strata 1 and 21 are collapsed and contain 20 *block-strings*. This collapsed stratum is targeted to be partitioned into two pseudo-strata. We randomly take 2 units from density stratum 21 and 8 units from density stratum 1 and place them into a pseudo-stratum 1a. The balance is placed in pseudo-stratum 1b.

Note that in this example, the 4 new pseudo-strata are of roughly the same size, i.e., they have approximately the same number of *block-strings* and can be thought of as having low, moderate or high concentrations of blacks or Hispanics.

Next, a mixing of *block-strings* is performed to form 2 pseudo-PSUs in each pseudo-density stratum.

M.1 If original density stratum g is in the new pseudo-stratum and n_g is even, then randomly assign $n_g/2$ string-units to pseudo-PSU 1 and the balance to pseudo-PSU 2.

M.2 If some of the original density strata have n_g odd, then use a randomization which assigns on the average $n_g/2$ *block-string* units to both

pseudo-PSUs 1 and 2 and also achieves approximate total pseudo-PSU balance. This balance may be achieved, for example, if one uses random and systematic assignments with random starts.

Table 3 demonstrates the mixing procedure as performed on the pseudo-strata of Table 2. Note that the original density strata components are now spread over 2 pseudo-PSUs and each pseudo-PSU covers multiple original strata. Such a design reduction should lessen identification risk. For additional discussion of the use of stratum mixing to reduce identification risk see Eltinge (1999). Now, a variance estimator for a pseudo-stratum total is

$$(\hat{Y}_{pseudo\text{-}PSU\ 1} - \hat{Y}_{pseudo\text{-}PSU\ 2})^2$$

If all the original n_g are even integers within a given pseudo-stratum, then the difference $(\hat{Y}_{pseudo\text{-}PSU\ 1} - \hat{Y}_{pseudo\text{-}PSU\ 2})$ has a mean equal to zero. Thus in this case, *stratum mixing* does not induce any additional variance estimator bias. This is in contrast with customary *stratum collapse* which tends to produce positively biased variance estimators.

An examination of Table 3 shows that the variance for pseudo strata 1a, and 1b should have mixing-bias equal to zero, but pseudo-strata 2a and 3a may have small positive-bias components. In these latter cases the step C.1 attempt to force a similarity among the grouped units should help to reduce collapsing bias. We anticipate that in most cases we can treat the variance estimator defined by this outlined strategy as having a small relative bias. The main drawback of this variance estimator is it loses degrees of freedom as compared to the original design.

Comment on NSR strata: If NSR strata exist, then the original two sampled PSUs can remain unmodified. If the new SR pseudo stratum weighted totals are of the same order of magnitude as the NSR strata, then the all the pseudo-PSUs will tend to have about the same magnitude in numbers of households and total weight as do the NSR PSUs. This characteristic makes the distinguishability between SR and NSR areas (frequently having metro and non-metro correspondence) more difficult. For some states this may help to reduce further the identifiability of the original selected sample units.

In practice the numbers of units to collapse should be chosen on a state-by-state basis.

Table 3: Mixing of *Block-Strings* to Form Pseudo-PSUs

block-string	original density-stratum	pseudo stratum	pseudo PSU
1	1	1a	1
2	1	1a	1
3	1	1a	1
4	1	1a	1
5	21	1a	1
6	1	1a	2
7	1	1a	2
8	1	1a	2
9	1	1a	2
10	21	1a	2
21	2	2a	1
22	2	2a	1
23	6	2a	1
24	6	2a	1
25	7	2a	1
26	2	2a	2
27	3	2a	2
28	6	2a	2
29	11	2a	2
30	7	2a	2
31	8	3a	1
32	14	3a	1
33	16	3a	1
34	16	3a	1
35	18	3a	1
36	19	3a	1
37	8	3a	2
38	9	3a	2
39	14	3a	2
40	16	3a	2
41	17	3a	2
42	18	3a	2

Table 4: Standard error ratios

variable / subdomains	$\sqrt{\frac{\hat{V}_{BRR}}{\hat{V}_{OD}}}$	$\sqrt{\frac{\hat{V}_{FC}}{\hat{V}_{OD}}}$
Number of persons diagnosed with hypertension		
Adults	0.96	1.05
Black Adults	1.24	1.33
Black Female Adults	1.09	1.20
Number of persons who have had an HIV test		
Adults	0.98	1.04
Black Adults	0.58	1.29
Black Female Adults	0.31	1.24
Mean BMI (Body Mass Index) (Weight in kg)/(Height in meters)^2		
Adults	0.98	1.00
Black Adults	1.13	1.11
Black Female Adults	1.30	1.16
Black 18-45 Adults	0.97	1.10

4 Applications to NHIS Data

To demonstrate the ideas above we will use the NHIS one-sample-adult-per-family subsample taken from a self-representing area of one state. For state-level variance estimation we consider three options

OD Original design with moderate collapse ("best"): $\hat{V}_{OD}$ with 196 *df*

FC Full collapse of all SR density strata within the entire state: $\hat{V}_{FC}$ with 225 *df*

BRR Balanced repeated replication estimator: $\hat{V}_{BRR}$ with 25 *df*

This was based upon a *2-pseudo-PSUs-per-pseudo-stratum* with the following characteristics:

i. Poststratification adjustments were carried out separately for each replicate sample using coarse Current Population Survey age-race-sex state controls.

ii. Fay adjustment with perturbation 70% were used (Fay (1984) and Judkins (1990)).

The variance estimator $\hat{V}_{FC}$ pools all *block-strings* into one large collapsed SR stratum. Removal of original strata may help avoid identification, but *block-strings* remain intact. Furthermore, if the state has a small NSR component, the size differential between SR and NSR units remains; and so identification risks may still remain. In addition, $\hat{V}_{FC}$ may suffer from significant collapse-bias. The proposed estimator $\hat{V}_{BRR}$ is structured to avoid identification risk, but at the expense of a loss of degrees of freedom.

Some comparisons of these variance estimators for population totals and means are presented in Table 4. There are several *caveats* to consider when making comparisons among these three estimators.

i. The variance estimator $\hat{V}_{OD}$ is computed with the poststratified weight treated as a sampling weight. Anecdotal comments from data users suggest that most users use fundamental sampling design structures but with a final sampling weight that incorporates all the weighting adjustments.

ii. From our limited study, it appears that poststratification will have substantially more impact at the state level than has typically been observed for national level estimation. Roughly speaking, at the state level there are sharp distinctions between samples and age-race-ethnicity-sex control totals. While distinctions are still present at the national levels, they appear to be smoothed out, due to effects of larger sample sizes.

iii. The nominal degrees of freedom terms are computed for a characteristic that is spread somewhat uniformly over the state. The black domains presented in Table 4 will not satisfy this uniformity condition and may have fewer degrees of freedom than nominal amount.

Comments on Table 4: Our study is quite limited, but the following observations are of interest. The estimator $\hat{V}_{FC}$ appears to suffer from bias due to collapsing of strata. Its magnitude is consistently larger than that of $\hat{V}_{OD}$. For all three estimators, the standard errors appear to be of the same magnitude on the adult domain which covers the full state. This may be evidence that the goal *(a)* is being met for domains covering the state.

For the black domains the actual sample size was about 180 persons, but the measured prevalence rates were at least 30% of the black population. Such characteristics will meet the NCHS publication standard that an estimator's coefficient of variation must not exceed 30%, and any estimator satisfying this requirement should be in scope of evaluation. In this situation the observed relation between $\hat{V}_{OD}$ and $\hat{V}_{BRR}$ did not show a consistent trend. A likely explanation is that black subpopulation is geographically concentrated and stability of the variance estimator is overestimated by the nominal 25 degrees of freedom.

5 Reduction of Degrees of freedom: Practical Effects on Inference

For our example the proposed variance estimator for reducing identification risk has only 25 nominal degrees of freedom as compared to the original design's nominal 196 degrees of freedom. One way in which to assess the practical impact of a reduction in degrees of freedom is to compare confidence intervals for a given population mean. For example, Table 5 presents nominal 95% confidence intervals for the mean body mass index (BMI) for the black adult population in the self-representing area of our specified state. These confidence intervals used the same customary sample ratio $\hat{\mu}$, but used the variance estimators $\hat{V}_{OD}$ and $\hat{V}_{BRR}$, respectively. More generally, one can use *p-value curves* to assess the practical impact of distinctions between variance estimation methods.

To develop this idea consider a typical 2-sided hypothesis test:
$H_0 : \mu = \mu_0$ vs $H_1 : \mu \neq \mu_0$
If the sample sizes are large, the analyst frequently assumes that the test statistic
$\hat{t}(\mu_0) = \frac{(\hat{\mu}-\mu_0)}{se(\hat{\mu})}$ has its *p-value* computed by

$$\hat{p}(\mu_0) = P[\,|t_d| > |\hat{t}(\mu_0)|\,\Big|\,\hat{t}(\mu_0)\,] \text{ where}$$

t_d is distributed as a central t random variable with d degrees of freedom.

A plot of $\hat{p}(\mu_0)$ against μ_0 gives the resulting *p-value curve*. In addition, projection of *p-value curve* onto horizontal axis at a specified $\hat{p}(\mu_0) = \alpha/2$ gives a test-inversion $(1-\alpha)100\%$ confidence interval.

Such a graphical display links statistical properties (width of confidence interval, slope of p-value curves) with practical significance (distinctions among competing μ_0 values of substantive interest). Some p-value curves were displayed during the presentation of this paper, but they are omitted here due to lack of space. Instead, Table 5 presents an example of a 95% confidence interval for a mean body mass index (BMI) for black adults.

Table 5: 95% Confidence Intervals for BMI

$\hat{V}$	$\hat{\mu}$	$se(\hat{\mu})$	df	lower	upper
OD	27.24	0.355	196	26.54	27.94
BRR	27.24	0.403	25	26.41	28.07

6 Summary

State-level NHIS analyses require special techniques to avoid micro-level geographical identification of the primary sample. This paper has proposed and evaluated an NHIS public-use approximate design for variance estimation. The key features are:

1. Self-representing areas are collapsed and partitioned.
2. A *2-pseudo-PSUs-per-pseudo-Stratum* design that can use standard BRR methods is targeted for construction.
3. Stratum mixing is implemented to reduce disclosure and identification risks.

7 References

Bureau of the Census (1977), The Current Population Survey: Design and Methodology, Technical Paper 40, U.S. Government Printing Office, Washington, D.C.

Eltinge, John L., (1999), Use of Stratum Mixing to Reduce Primary-Unit-Level Identification Risk in Public-Use Survey Datasets. *Proceedings of the 1999 Federal Committee on Statistical Methodology Research Conference*, to appear.

Fay, Robert E., (1984), Some properties of estimates of variance based on replication methods, *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 495- 500.

Judkins, David R., (1990), Fay's method for variance estimation, *Journal of Official Statistics*, 6, 223-239.

National Center for Health Statistics (1999), National Health Interview Survey: Research for the 1995 2004 redesign, *Vital and Health Statistics*, Series 2, No. 126.

National Center for Health Statistics (2000), National Health Interview Survey: Design and Estimation for the National Health Interview Survey, 1995-2004, *Vital and Health Statistics*, Series 2, to appear.

Wolter, K. M.,(1985), *Introduction to Variance Estimation* Springer:Berlin:New York.